

### **REMARKS/ARGUMENTS**

Claims 28-35 and 38-40 are pending in this application.

Claims 28-32 have been amended to further clarify what applicants have always regarded as their invention. Support for polypeptides "comprising" polypeptide variants is found in the specification at, for example, page 283, lines 2-27. The amendments to the claims are fully supported by the specification and claims as originally filed and do not constitute new matter. Applicants believe that the current amendments place all claims in *prima facie* condition for allowance or, at least, in a better form for consideration on appeal. Accordingly, the consideration and entry of the present amendment after final rejection is respectfully requested.

Applicants expressly reserve the right to pursue any canceled matter in subsequent continuation, divisional or continuation-in-part application(s).

In addition, Applicants request the PTO to take note of the Revocation and Power of Attorney and Change of Address filed on February 28, 2003, and kindly direct all future correspondence to the address indicated, *i.e.*, to:

**CUSTOMER NO. 35489**  
**Ginger R. Dreger**  
**HELLER EHRMAN LLP**  
**275 Middlefield Road**  
**Menlo Park, California 94025**  
**Telephone: (650) 324-7000**  
**Facsimile: (650) 324-0638**

**I. Claim Rejections Under 35 U.S.C. §101 and 35 U.S.C. §112, First Paragraph**

Claims 28-35 and 38-40 remain rejected under 35 U.S.C. §101 allegedly "because the claimed invention is not supported by either a credible, specific and substantial asserted utility or a well established utility." (Page 2 of the instant Office Action).

Claims 28-35 and 38-40 remain further rejected under 35 U.S.C. §112, first paragraph, allegedly "since the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention." (Page 19 of the instant Office Action).

For the reasons outlined below, Applicants respectfully disagree and traverse the rejection. With respect to Claims 28-35 and 38-40, Applicants submit that not only has the Patent Office not established a *prima facie* case for lack of utility and enablement, but that the PRO1759 polypeptides possess a credible, specific and substantial asserted utility and are fully enabled.

**The data and documentary evidence supporting a patentable utility**

First of all, Applicants respectfully maintain the position that the specification discloses at least one credible, substantial and specific asserted utility for the claimed PRO1759 polypeptides for the reasons previously set forth in Applicants' Responses filed on February 2, 2005, and July 21, 2005.

Furthermore, as first discussed in Applicants' Response of February 2, 2005, Applicants rely on the gene amplification data for patentable utility of the PRO1759 polypeptide, and the gene amplification data for the gene encoding the PRO1759 polypeptide is clearly disclosed in the instant specification under Example 143. A  $\Delta C_t$  value of at least 1.0 was observed for PRO1759 in at least three of the tumors listed in Table 8. PRO1759 showed approximately 1.11-1.51  $\Delta C_t$  units which corresponds to  $2^{1.11}$ - $2^{1.51}$  fold amplification or 2.16 fold to 2.85-fold amplification in lung tumors HF000842 and HF001296, and in colon tumor center HF000795. (See Table 8 of the specification). Accordingly, the present specification clearly discloses overwhelming evidence that the gene encoding the PRO1759 polypeptide is significantly amplified in lung and colon tumors.

The Examiner has asserted that "the instant specification provides data showing a very small increase in DNA copy number" (Page 3 of the instant Office Action). The Examiner has further asserted that "[t]he specification merely demonstrates that the PRO1759 nucleic acid was amplified in three cancer samples, to a minor degree (about 2.5 fold)." (Page 13 of the instant Office Action).

Applicants submit that the Examiner seems to have applied a heightened utility standard in this instance, which is legally incorrect. Applicants have shown that the gene encoding PRO1759 demonstrated significant amplification, from 2.16 to 2.85 fold, in three lung and colon

tumors. As explained in the Declaration of Dr. Audrey Goddard (submitted with the Response filed February 2, 2005):

It is further my considered scientific opinion that an at least **2-fold increase** in gene copy number in a tumor tissue sample relative to a normal (*i.e.*, non-tumor) sample **is significant** and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. (Emphasis added).

By referring to the 2.16-fold to 2.85-fold amplification of the PRO1759 gene in lung tumors as "very small," or "minor," the Examiner appears to ignore the teachings within an expert's declaration without any basis, or without presenting any evidence to the contrary. Appellants respectfully draw the Board's attention to the Utility Examination Guidelines (Part IIB, 66 Fed. Reg. 1098 (2001)) which state that:

"Office personnel must accept an opinion from a qualified expert that is based upon relevant facts whose accuracy is not being questioned; it is improper to disregard the opinion solely because of a disagreement over the significance or meaning of the facts offered".

Thus, given the absence of any evidence to the contrary, Applicants maintain that the 2.16 to 2.85-fold amplification disclosed for the PRO1759 gene is significant and forms the basis for the utility claimed herein.

Further, Applicants respectfully submit that the amplification of the nucleic acids in even one lung or colon tumor provides specific and substantial utility for PRO1759 as a diagnostic marker of the type of lung or colon tumor in which the gene encoding it was amplified.

Applicants submit that the tumors listed in Table 8 are not similar tumors from different patients, but various types/classes of lung and/or colon tumors at different stages. Accordingly, a positive result from one tumor, where the nucleic acid was amplified, but not from other tumors, indicates that PRO1759 can be used as a marker for diagnosing the presence of that kind of tumor in which it was amplified. Amplification of the nucleic acid would be indicative of that specific class of lung or colon tumor, whereas absence of amplification would be non-conclusive.

The Examiner has asserted that "[o]ne cannot determine from the data in the specification whether the observed 'amplification' of nucleic acid is due to increase in chromosomal copy

number, or alternatively due to an increase in transcription rates." (Page 13 of the instant Office Action). Applicants note that the data in the specification relates to amplification of DNA, not mRNA, thus transcription rates would not affect this data.

The Examiner further asserts that "the data were not corrected for aneuploidy, a phenomenon that occurs in cancerous and non-cancerous lung (Sen). Therefore, it is not clear that the reported amplification is significant." (Page 13 of the instant Office Action). No evidence is cited in support of the assertion that aneuploid DNA is found in normal tissue. To the contrary, Sen, the only reference cited by the Examiner, teaches that aneuploidy is "a discrete chromosome mutation event that contributes to malignant transformation and progression process" (Abstract). Sen further teaches that "[i]n addition to being implicated in tumorigenesis and correlated with distinct tumor phenotypes, chromosome aneuploidy has been used as a marker of risk assessment and prognosis in several other cancers" (page 84, col. 1). Therefore, aneuploidy is a useful marker for a cancerous or pre-cancerous state. Hence, Applicants respectfully submit that whether a pre-cancerous or tumor sample were analyzed, the showing of increased PRO1759 DNA levels as a result of aneuploidy would still be significant, since it would lead to the diagnosis of either a pre-cancerous state or a cancerous state, which is the utility asserted here.

Applicants further submit that it is known in the art that detection of gene amplification can be used for cancer diagnosis regardless of whether the increase in gene copy number results from intrachromosomal changes or from chromosomal aneuploidy. As explained by Dr. Ashkenazi in his Declaration (submitted with Applicants' Response filed February 2, 2005),

An increase in gene copy number can result not only from intrachromosomal changes but also from chromosomal aneuploidy. It is important to understand that detection of gene amplification can be used for cancer diagnosis even if the determination includes measurement of chromosomal aneuploidy. Indeed, as long as a significant difference relative to normal tissue is detected, it is irrelevant if the signal originates from an increase in the number of gene copies per chromosome and/or an abnormal number of chromosomes.

Hence, Applicants submit that gene amplification of a gene, whether by aneuploidy or any other mechanism, is useful as a diagnostic marker.

The Examiner further asserts that "[t]he control DNA appeared to be from blood rather than from a matched tissue sample (i.e., healthy lung and colon), while the literature shows that matched tissue samples are the standard (Pennica *et al.*).\" (Page 13 of the instant Office Action).

Applicants respectfully submit that the Examiner's statement is scientifically and factually inaccurate, because the negative control taught in the specification was known in the art at the time of filing, and accepted as a true negative control as demonstrated by use in peer reviewed publications, including Pennica *et al.* Pennica *et al.* explain that "[t]he relative WISP gene copy number in each colon tumor DNA was compared with **pooled normal DNA** from 10 donors by quantitative PCR\" (page 14720, col. 2; emphasis added). Pennica *et al.* further explain that DNA was isolated from \"the pooled blood of 10 normal human donors.\" (page 14718, col. 1). Thus Pennica *et al.* used the same control for their gene amplification experiments as that described in the instant specification.

In further examples, Pitti *et al.* (Exhibit F submitted with the Response filed February 2, 2005), used the same quantitative TaqMan PCR assay described in the specification to study gene amplification in lung and colon cancer of DcR3, a decoy receptor for Fas ligand. As described, Pitti *et al.* analyzed DNA copy number \"in genomic DNA from 35 primary lung and colon tumours, relative to pooled genomic DNA from peripheral blood leukocytes (PBL) of 10 healthy donors.\" (Page 701, col. 1; emphasis added). The authors also analyzed mRNA expression of DcR3 in primary tumor tissue sections and found tumor-specific expression, confirming the finding of frequent amplification in tumors, and confirming that the pooled blood sample was a valid negative control for the gene amplification experiments. In Bieche *et al.* (Exhibit G submitted with the Response filed February 2, 2005), the authors used the quantitative TaqMan PCR assay to study gene amplification of *myc*, *ccnd1* and *erbB2* in breast tumors. As their negative control, Bieche *et al.* used normal leukocyte DNA derived from a small subset of the breast cancer patients (page 663). The authors note that \"[t]he results of this study are consistent with those reported in the literature\" (page 664, col. 2), thus confirming the validity of the negative control. Accordingly, the art demonstrates that pooled normal blood samples are considered to be a valid negative control for gene amplification experiments of the type described in the specification.

**A prima facie case of lack of utility has not been established**

The Examiner asserts that "the specification provides data showing a very small increase in DNA copy number in two different types of tumor tissue (lung and colon)...However, there is no evidence regarding whether or not PRO1759 mRNA or polypeptide levels are also increased in these cancers." The Examiner further asserts that "what is often seen is a lack of correlation between DNA amplification and increased peptide levels." (Page 4 of the instant Office Action).

In support of these assertions, the Examiner refers to previously cited papers by Pennica *et al.*, Haynes *et al.*, Chen *et al.* and Hu *et al.* The Examiner also cites additional papers by Madoz-Gurpide *et al.*, Celis *et al.*, Steiner *et al.* and Feroze-Merzoug *et al.* in support of the assertion that "the state of the art is such that polypeptide levels cannot be accurately predicted from mRNA levels." (Page 5 of the instant Office Action).

Applicants respectfully remind the Examiner that the evidentiary standard to be used throughout *ex parte* examination of a patent application is a preponderance of the totality of the evidence under consideration. Accordingly, Applicants submit that in order to overcome the presumption of truth that an assertion of utility by the applicant enjoys, the Examiner must establish that **it is more likely than not** that one of ordinary skill in the art would doubt the truth of the statement of utility. Therefore, it is not legally required that there be a "necessary" correlation between the data presented and the claimed subject matter, such that the amplification of every possible gene inevitably results in protein over-expression. The law requires only that one skilled in the art should accept that such a correlation is **more likely than not to exist**. Applicants respectfully submit that when the proper evidentiary standard is applied, a correlation must be acknowledged.

Applicants respectfully submit that, for the reasons previously set forth in the Applicants' response filed on February 2, 2005, the teachings of Pennica *et al.* are specific to *WISP* genes, and say nothing about the correlation of gene amplification and protein expression in general.

The Examiner reiterates that Haynes *et al.* state that "[p]rotein expression levels are not predictable from the RNA expression levels." (Page 7 of the instant Office Action). As discussed in Applicants' previous Responses, Haynes *et al.* teaches that "there was a *general trend but no strong correlation* between protein [expression] and transcript levels" (Emphasis

added). In Figure 1, there is a positive correlation between mRNA and protein amongst *most* of the 80 yeast proteins studied. Thus Haynes *et al.* support Applicants' position that there is a general correlation between mRNA and protein expression levels.

Haynes *et al.* may teach that protein levels cannot be "accurately predicted" from mRNA levels in the sense that the exact numerical amounts of protein present in a tissue cannot be determined based upon mRNA levels. Applicants respectfully submit that the Office Action's emphasis on the need to "accurately predict" protein levels based on mRNA levels misses the point. The asserted utility for the claimed polypeptides is in the diagnosis of cancer. What is relevant to use as a cancer diagnostic is relative levels of gene or protein expression, not absolute values, that is, that the gene or protein is differentially expressed in tumors as compared to normal tissues. Applicants need only show that there is a correlation between DNA, mRNA, and protein levels, such that gene amplification and mRNA overexpression generally predict protein overexpression. A showing that mRNA levels can be used to "accurately predict" the precise levels of protein expression is not required.

The Examiner further cites Feroze-Merzoug *et al.*, to the effect that a lack of correlation between mRNA and protein is evident even in yeast. These conclusions are based upon a citation to a paper by Gygi *et al.* (Mol. Cell. Biol. 19:1720-1730 (1999); copy enclosed as Exhibit 1). Gygi *et al.*, however, clearly teach that "there was a general trend of increased protein levels resulting from increased mRNA levels." (Emphasis added. See page 1726, left column, second paragraph and Figure 5). Gygi *et al.* further states that the "correlation coefficient for the whole data set (106 genes) was 0.935." In addition, Gygi *et al.* state that the correlation coefficient for genes where the message level was below 10 copies/cell was 0.356 and for most abundant proteins was 0.94. (See page 1726, second paragraph and page 1727, left column, second paragraph). Therefore, Gygi *et al.* disclose positive correlations for both lower abundance and higher abundance proteins. Gygi *et al.* clearly teach that there is a general trend of increased protein levels from increased mRNA levels for all levels of mRNA expression. Thus, again, Feroze-Merzoug *et al.* appear to be focusing on "accurately predicting" the precise levels of protein expression, which is not required for utility as a cancer diagnostic.

Applicants further note that Feroze-Merzoug *et al.* looked specifically at androgen regulated genes, which were not necessarily associated with cancer. The expression of these genes clearly involves different biological process than in lung or colon tumor development. Therefore, even if the teaching of Feroze-Merzoug *et al.* accurately reflects the correlation between mRNA and protein for the particular system studied, it does not apply to the lung and colon cancer diagnostic assays of the present application.

The reference by Chen *et al.*, as discussed in Applicants' Response filed July 22, 2005 is not applicable to the present application, in part because the Chen paper does not account for different expression in different tissues or different stages of cancer. Further, a review of the actual data presented in Tables I and II of the Chen paper demonstrates that it is more likely than not that protein levels will correlate with mRNA expression levels. Once again, the Examiner focuses on statements in the Chen paper to the effect that protein levels cannot be "predicted" from mRNA levels, and that mRNA expression patterns are "insufficient for understanding the expression of protein products." (Page 8 of the instant Office Action). As discussed above, there is no requirement that protein levels be "accurately predicted."

The same authors in Chen *et al.*, published a later paper, Beer *et al.*, which described gene expression of genes in adenocarcinomas and compared that to protein expression. The results demonstrate that the authors of the Chen paper agree that microarrays provide a reliable measure of the expression levels of the gene and can be used to identify genes whose overexpression is associated with tumors. The Examiner asserts that Beer *et al.* is not relevant to the instant application because the authors also did additional northern blot and immunohistochemistry experiments for three of their identified genes. The Examiner asserts that "the specification of the instant application does not complement the low (2-fold) PRO1759 gene expression data with any mRNA or protein studies." (Pages 9-10 of the instant Office Action). In support of the role of proteomics as a "complementary technology" to DNA microarrays, the Examiner cites references by Madoz-Gurpide *et al.*, Celis *et al.* and Steiner *et al.*, in addition to Feroze-Merzoug *et al.*

Applicants respectfully point out that Madoz Gurpide *et al.* state only that it is "unclear" how well RNA levels reported correlate with protein levels. In support of this assertion, the



authors cite only a single reference, the same paper by Chen *et al.* previously cited by the Examiner, which Applicants have already discussed. Madoz Gulpide *et al.* also acknowledge that DNA microarray studies, such as those carried out by Beer *et al.*, specifically cited by the authors at page 52, "justify the use of this technology for uncovering patterns of gene expression that are clinically informative" (page 53).

Applicants respectfully submit that while proteomics is indeed a complementary technology to DNA microarrays, this does not mean that proteomic experiments are required in addition to measurements of mRNA levels to determine protein expression. The cited references make clear that proteomic techniques are useful to obtain information beyond expression levels, such as the protein's activation state, posttranslational modifications, and subcellular localization. For example, Madoz-Gulpide *et al.* explain that mRNA expression alone does not provide information regarding "activation state, post-translational modification or localization of corresponding proteins" (page 168, col. 1). Celis *et al.* note that "proteomics addresses problems that cannot be approached by DNA analysis, namely, relative abundance of the protein product, post-translational modification, subcellular localization, turnover, interaction with other proteins as well as functional aspects" (page 6, col. 2).

While this additional information may be useful in elucidating the detailed biological function of a protein, it is not required to establish utility of a protein as a marker for cancer. The claimed PRO1759 polypeptides can be used in cancer diagnosis without any knowledge regarding the function or cellular role of the polypeptides. Applicants submit that the law clearly states that "it is not a requirement of patentability that an inventor correctly set forth, or even know, how or why the invention works." *Newman v. Quigg*, 11 U.S.P.Q.2d 1340 (Fed. Cir. 1989). Accordingly, the disclosure or identification of the mechanism by which PRO1759 is associated with cancer is not required in order to establish the patentable utility of the claimed PRO1759 polypeptides. Thus while Madoz-Gulpide *et al.* note that it is "more difficult to develop an understanding of disease at a mechanistic level using DNA microarrays," (page 53) this is not relevant to Applicants' assertions of utility, since, as discussed above, it is not necessary to understand how or why an invention works in order to demonstrate utility.

Applicants note that the cited references make clear that proteomic data is useful primarily in the areas of drug development, in order to better understand the functional mechanisms of the protein target. These considerations are not relevant to use as a cancer diagnostic. Thus Steiner *et al.* does not even discuss cancer or diagnostics, but is limited to the role of proteomics in drug development and toxicology testing. Celis *et al.* explain that proteomics is useful in combination with arrays "for the entire process of drug development and evaluation." (page 6; col. 1).

Applicants further submit that significant correlations between gene and protein expression are most likely to be observed for genes associated with cancer, since as Celis *et al.* note, "transformation resulted in the abnormal expression of normal genes, rather than in the expression of new ones" (page 11, col. 1). Accordingly, alterations in gene amplification or expression are more likely to be associated with altered protein expression in the case of cancer than in other cases where DNA microarrays are used (for example, the study of androgen-regulated genes described in Feroze-Merzoug *et al.*), because, as explained by Celis *et al.*, the alterations in expression levels of certain normal proteins are part of the process that leads to cancer.

In their discussion of DNA microarrays and proteomics applied to the same samples, Celis *et al.* cite Orntoft *et al.*, and note that "in most cases there was a good correlation between transcript and protein levels." Celis *et al.* further explain that those few cases which showed apparent discrepancies may have been due to other causes, such as post-transcriptional processing or degradation of the protein, or the choice of methods used to assess protein expression levels. Celis *et al.* also note that the observation that there is often more change in mRNAs as compared to the proteins may be due to the fact that current technologies detect mainly high abundance proteins, while most of the changes affecting protein levels may involve low abundance proteins. Thus the correlation between mRNA and protein levels may be even higher than typically observed, given these factors.

Finally, the Examiner again cites Hu *et al.* in support of the assertion that mRNA expression is not correlated with protein expression. Applicants respectfully submit that Hu *et al.* does not conclusively show that it is more likely than not that gene amplification does not

result in increased expression at the mRNA and polypeptide levels. Applicants respectfully submit that Hu *et al.* manipulated various aspects of the input data in order to minimize the false positives and negatives in their analysis. Applicants further submit that the statistical analysis by Hu *et al.* is not a reliable standard because the frequency of citation only reflects the current research interest in a molecule but not the true biological function of the molecule. Finally, the conclusion in Hu *et al.* only applies to a specific type of breast tumor (estrogen receptor (ER)-positive breast tumor) and can not be generalized as a principle governing microarray study of breast cancer in general, let alone the various other types of cancer genes in general. In fact, even Hu *et al.* admit that "[i]t is likely that this threshold will change depending on the disease as well as the experiment. Interestingly, the observed correlation was only found among ER-positive (breast) tumors not ER-negative tumors." (See page 412, left column). Therefore, based on these findings, the authors add, "This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently." (*Id.*; emphasis added).

The Examiner asserts that "Applicant is holding Hu *et al.* to a higher standard than their own specification, which does not provide proper statistical analysis such as reproducibility, standard error rates, etc." (Page 11 of the instant Office Action). Applicants note that they do not argue that Hu *et al.* lacks reproducibility, standard error rates, etc. for their data, given that Hu *et al.* did a literature survey and conducted no actual experiments of their own. Rather, Applicants' point is that, given the various biases in selecting the data to be considered, as acknowledged by the authors themselves, the collection of data surveyed by Hu *et al.* simply does not demonstrate the conclusion the PTO attempts to reach concerning a general lack of correlation between microarray data and biological significance. Accordingly, Applicants respectfully submit that the Examiner has not shown a lack of correlation between microarray data and the biological significance of cancer genes.

The Patent Office has failed to meet its initial burden of proof that Applicant's claims of utility are not substantial or credible. The arguments presented by the Examiner in combination with the Pennica, Haynes, Hu, and Chen papers, as well as the Madoz-Gurpide, Steiner, Celis

and Feroze-Merzoug papers, do not provide sufficient reasons to doubt the statements by Applicants that PRO1759 has utility. As discussed above, the law does not require the existence of a "necessary" correlation between mRNA and protein levels. Nor does the law require that protein levels be "accurately predicted." According to the authors themselves, the data in the above cited references confirm that there is a general trend between protein expression and transcript levels, which meets the "more likely than not standard" and show that a positive correlation exists between mRNA and protein. Therefore, Applicants submit that the Examiner's reasoning is based on a misrepresentation of the scientific data presented in the above cited reference and application of an improper, heightened legal standard. In fact, contrary to what the Examiner contends, the art indicates that, if a gene is amplified in cancer, it is more likely than not that the encoded protein will be expressed at an elevated level.

**It is "more likely than not" for amplified genes to have increased mRNA and protein levels**

Applicants have submitted ample evidence to show that, in general, if a gene is amplified in cancer, it is more likely than not that the encoded protein will be expressed at an elevated level. First, the articles by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.*, (made of record in Applicants' Response filed February 2, 2005) collectively teach that in general, gene amplification increases mRNA expression. Second, the Declaration of Dr. Paul Polakis, principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, shows that, in general, there is a correlation between mRNA levels and polypeptide levels. Thus, taken together, all of the submitted evidence supports Appellants' position that gene amplification is more likely than not predictive of increased mRNA and polypeptide levels.

The Examiner first asserts that "Hanna *et al.* clearly show that the skilled artisan does not assume that any tumor with a HER-2/neu gene amplification event also overexpressed HER-2/neu protein. It is tested empirically." (Page 14 of the instant Office Action). Applicants respectfully point out that the Examiner appears to have misread Hanna *et al.* Hanna *et al.* clearly state that gene amplification (as measured by FISH) and polypeptide expression (as

measured by immunohistochemistry, IHC) are well correlated ("in general, FISH and IHC results correlate well" (page 1, col. 2)). It is only a subset of tumors which show discordant results. Thus Hanna *et al.* support Applicants' position that it is more likely than not that gene amplification correlates with increased polypeptide expression.

Applicants also note that Hanna *et al.* was not cited specifically as evidence that gene amplification and protein expression are well correlated, but as evidence that even if gene amplification does not result in overexpression of the gene product (*i.e.*, the protein) an analysis of the expression of the protein is useful in determining the course of treatment. The Examiner appears to view the testing described in the Ashkenazi Declaration and the Hanna paper as experiments involving further characterization of the PRO1759 polypeptide itself. In fact, such testing is for the purpose of characterizing not the PRO1759 polypeptide, but the tumors in which the gene encoding PRO1759 is amplified. The PRO1759 polypeptide is therefore useful in tumor categorization, the results of which become an important tool in the hands of a physician enabling the selection of a treatment modality that holds the most promise for the successful treatment of a patient.

The Examiner further asserts that "Orntoft *et al.* could only compare the levels of about 40 well-resolved and focused *abundant* proteins." (Page 16 of the instant Office Action; emphasis in original). While technical considerations did prevent Orntoft *et al.* from evaluating a larger number of proteins, the ones they did look at showed a clear correlation between mRNA and protein expression levels. As Orntoft *et al.* state, "In general there was a highly significant correlation ( $p < 0.005$ ) between mRNA and protein alterations.... 26 well focused proteins whose genes had a known chromosomal location were detected in TCCs 733 and 335, and of these 19 correlated ( $p < 0.005$ ) with the mRNA changes detected using the arrays." (See page 42, column 2 to page 34, column 2). Accordingly, Orntoft *et al.* clearly support Applicants' position that proteins expressed by genes that are amplified in tumors are useful as cancer markers. Applicants further note that while Orntoft *et al.* did not compare cancerous versus non-cancerous tissues, they did compare invasive versus benign tumors, thus finding genes that were markers of tumor malignancy.

The Examiner also appears to misunderstand the data presented by Hyman *et al.* The Examiner has asserted that "of the 12,000 transcripts analyzed, a set of 270 was identified in which overexpression was attributable to gene amplification." The Examiner concludes that "[t]his proportion is approximately 2%; the Examiner maintains that 2% does not provide a reasonable expectation that the slight amplification of PRO351 would be correlated with elevated levels of mRNA, much less protein." (Page 17 of the instant Office Action). Applicants respectfully submit that the Examiner appears to have misinterpreted the results of Hyman *et al.* Hyman *et al.* chose to do a genome-wide analysis of a large number of genes, most of which, as shown in Figure 2, were not amplified. Accordingly, the 2% number is meaningless, as the low figure mainly results from the fact that only a small percentage of genes are amplified in the first place. The significant figure is not the percentage of genes in the genome that show amplification, but the percentage of amplified genes that demonstrate increased mRNA and protein expression.

The Examiner has further asserted that the Hyman reference "found 44% of *highly* amplified genes showing overexpression at the mRNA level, and 10.5% of *highly* overexpressed genes being amplified; thus, even at the level of high amplification and high overexpression, the two do not correlate." (Page 17 of the instant Office Action). Applicants submit that the 10.5% figure is not relevant to the issue at hand. One of skill in the art would understand that there can be more than one cause of overexpression. The issue is not whether overexpression is always, or even typically caused by gene amplification, but rather, whether gene amplification typically leads to overexpression.

The Examiner's assertion is not consistent with the interpretation Hyman *et al.* themselves place on their data, stating that, "The results illustrate **a considerable influence of copy number on gene expression patterns.**" (page 6242. col. 1; emphasis added). In the more detailed discussion of their results, Hyman *et al.* teach that "[u]p to 44% of the highly amplified transcripts (CGH ratio, >2.5) were overexpressed (*i.e.*, **belonged to the global upper 7% of expression ratios**) compared with only 6% for genes with normal copy number." (See page 6242, col. 1; emphasis added). These details make it clear that Hyman *et al.* set a highly restrictive standard for considering a gene to be overexpressed; yet almost half of all highly

amplified transcripts met even this highly restrictive standard. Therefore, the analysis performed by Hyman *et al.* clearly shows that "it is more likely than not" that a gene which is amplified in tumor cells will have increased gene expression.

The Examiner asserts that Orntoft *et al.*, Hyman *et al.* and Pollack *et al.* did not study lung or colon cancer. Applicants respectfully note that only one of the references cited by the Examiner, Chen *et al.*, was related to either lung or colon cancer. As discussed above, Tables I and II of the Chen paper demonstrates that it is more likely than not that protein levels will correlate with mRNA expression levels.

The Examiner further asserts that Hyman *et al.* and Pollack *et al.* do not examine protein expression. Applicants respectfully submit that the Orntoft *et al.*, Hyman *et al.* and Pollack *et al.* references were submitted primarily as evidence that in general, gene amplification increases mRNA expression. With regard to the correlation between mRNA expression and protein levels, Applicants previously submitted a Declaration by Dr. Polakis, principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, to show that mRNA expression correlates well with protein levels, in general.

Applicants submit that Dr. Polakis' Declaration is presented to support the position that there is a correlation between mRNA levels and polypeptide levels. Applicants emphasize that the opinions expressed in the Polakis Declaration are all based on factual findings. Thus, Dr. Polakis explains that in the course of their research using microarray analysis, he and his co-workers identified approximately 200 gene transcripts that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. Subsequently, antibodies binding to about 30 of these tumor antigens were prepared, and mRNA and protein levels were compared. In approximately 80% of the cases, the researchers found that increases in the level of a particular mRNA correlated with changes in the level of protein expressed from that mRNA when human tumor cells are compared with their corresponding normal cells. Dr. Polakis' statement that "an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell" is based on factual, experimental findings, clearly set forth in the Declaration. Accordingly, the Declaration is not merely conclusive, and the fact-based conclusions of Dr.

Polakis would be considered reasonable and accurate by one skilled in the art. As previously discussed, the Utility Examination Guidelines<sup>1</sup> state, "Office personnel must accept an opinion from a qualified expert that is based upon relevant facts whose accuracy is not being questioned; it is improper to disregard the opinion solely because of a disagreement over the significance or meaning of the facts offered."

Taken together, although there are some examples in the scientific art that do not fit within the central dogma of molecular biology that there is a correlation between polypeptide and mRNA levels, these instances are exceptions rather than the rule. In the majority of amplified genes, the teachings in the art, as exemplified by Orntoft *et al.*, Hyman *et al.*, Pollack *et al.*, and the Polakis Declaration, overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO1759 gene, that the PRO1759 polypeptide is concomitantly overexpressed. Thus, Applicants submit that the claimed PRO1759 polypeptides have utility in the diagnosis of cancer and based on such a utility, one of skill in the art would know exactly how to use the claimed polypeptides for diagnosis of cancer.

Accordingly, Applicants respectfully request the Examiner to reconsider and withdraw the rejections of Claims 28-35 and 38-40 under 35 U.S.C. §101 and 35 U.S.C. §112, first paragraph.

### **III. Claim Rejections Under 35 U.S.C. §112, First Paragraph (Enablement)**

Claims 28-32 and 39-40 additionally remain rejected under 35 U.S.C. §112, first paragraph as allegedly lacking enablement for the claimed variant polypeptides having at least 80-99% identity to SEQ ID NO:374, wherein the nucleic acid encoding the polypeptide is amplified in lung or colon tumors.

Applicants have provided native PRO1759 sequence SEQ ID NO:374. The present application also describes methods for identifying polynucleotides which are amplified in lung or colon tumors. Example 143 of the present application provides detailed protocols and assays for gene amplification in lung and colon tumors. By following the disclosure in the specification,

---

<sup>1</sup> Part IIB, 66 Fed. Reg. 1098 (2001).



one skilled in the art can easily test whether the gene encoding a variant native sequence PRO1759 polypeptide is amplified in lung or colon tumors. The specification further describes methods for the determination of percent identity between two amino acid sequences. (See page 302, line 4 to page 305, line 4). In fact, the specification teaches specific parameters to be associated with the term "percent identity" as applied to the present invention. Accordingly, one of skill in the art could identify whether the variant PRO1759 native sequence falls within the parameters of the claimed invention. Once such an amino acid sequence was identified, the specification sets forth methods for making the amino acid sequences (see page 354, line 30 to page 358, line 34) and methods of preparing the PRO polypeptides (see page 358, line 35 and onward).

Therefore, Applicants respectfully submit that one of skill in the art could readily test a variant native sequence polypeptide to determine whether the polynucleotide encoding it is amplified in lung or colon tumors by the methods set forth in Example 143. Furthermore, one of ordinary skill in the art has a sufficiently high level of technical competence to identify sequences with at least 80% identity to SEQ ID NO:374. Accordingly, one of ordinary skill could practice the claimed invention without undue experimentation.

The Examiner asserts that "[n]o such variants were made or shown to have activity." (Page 21 of the instant Office Action). As discussed in the M.P.E.P. § 2164.08, "[t]he specification need not contain an example if the invention is otherwise disclosed in such manner that one skilled in the art will be able to practice it without an undue amount of experimentation. *In re Borkowski*, 422 F.2d 904, 908, 164 USPQ 642, 645 (CCPA 1970)." Given that, as discussed above, one of ordinary skill in the art could make and use the claimed variant sequences without any undue experimentation, there is no requirement that the specification provide examples of such variant polypeptides.

The Examiner asserts that "Applicant has provided little or no guidance beyond the mere presentation of sequence data to enable one of ordinary skill in the art to determine, without undue experimentation, the positions in the DNA and protein which are tolerant to change." (Page 21 of the instant Office Action). To the contrary, the specification provides detailed guidance as to changes that may be made to a PRO polypeptide without adversely affecting its

activity (page 354, line 30 to page 357, line 7). This guidance includes a listing of exemplary and preferred substitutions for each of the twenty naturally occurring amino acids (Table 6, page 356). Further, it is the activity of the encoding polynucleotide, not the polypeptide, that is recited in the claims. Thus considerations of sequences in the polypeptide sequence that are "critical to the protein's structure function relationship" are not relevant.

The claims currently recite polypeptide sequences associated with a biological activity of the encoding polynucleotides. This biological activity together with the well defined relatively high degree of sequence identity and general knowledge in the art at the time the invention was made, sufficiently defines the claimed genus such that, one skilled in the art, at the effective date of the present application, would have known how to make and use the claimed polypeptide sequences without undue experimentation. As the M.P.E.P. states, "[t]he fact that experimentation may be complex does not necessarily make it undue, if the art typically engages in such experimentation."<sup>2</sup>

As discussed above, a considerable amount of experimentation is permissible, if it is merely routine. Applicants submit that the identification of variant native sequence PRO1759 polypeptides having at least 80% identity to SEQ ID NO:374 wherein the polynucleotide encoding the polypeptide is amplified in lung or colon tumors, can be performed by techniques that were well known in the art at the priority date of this application, and that the performance of such work does not require undue experimentation.

For the above-noted reasons, Applicants respectfully request the Examiner to reconsider and withdraw the enablement rejection of Claims 28-32 and 39-40 under 35 U.S.C. §112, first paragraph.

#### **IV. Claim Rejections Under 35 U.S.C. §112, First Paragraph (Written Description)**

Claims 28-32 and 38-40 remain rejected under 35 USC 112, first paragraph, as allegedly lacking adequate written description for the claimed variant polypeptides having at least 80-99%

---

<sup>2</sup> M.P.E.P. §2164.01 citing *In re Certain Limited-charge Cell Culture Microcarriers*, 221 U.S.P.Q. 1165, 1174 (Int'l Trade Comm'n 1983), *aff' sub nom. Massachusetts Institute of Technology v A.B. Fortia*, 774 F 2d 1104, 227 U.S.P.Q. 428 (Fed. Cir. 1985).

identity to SEQ ID NO:374, wherein the nucleic acid encoding the polypeptide is amplified in lung or colon tumors.

Applicants respectfully submit that the instant specification evidences the actual reduction to practice of a full-length PRO1759 polypeptide of SEQ ID NO:374, with or without its signal peptide sequence. The Examiner has previously acknowledged that a polypeptide comprising the sequence set forth in SEQ ID NO:354 meets the written description provision of 35 U.S.C. §112, first paragraph. Thus, the genus of native sequence polypeptides with at least 80% sequence identity to SEQ ID NO:374, which possess the functional property that the nucleic acid encoding the polypeptide is amplified in lung and colon tumors, would meet the requirement of 35 U.S.C. §112, first paragraph, as providing adequate written description.

The specification describes methods for the determination of percent identity between two amino acid sequences. (See page 302, line 4 to page 305, line 4). In fact, the specification teaches specific parameters to be associated with the term "percent identity" as applied to the present invention. The specification further provides detailed guidance as to changes that may be made to a PRO polypeptide without adversely affecting its activity (page 354, line 30 to page 357, line 7). This guidance includes a listing of exemplary and preferred substitutions for each of the twenty naturally occurring amino acids (Table 6, page 356). The specification describes methods for one of ordinary skill in the art to identify polypeptides having at least 80% identity to SEQ ID NO:374 wherein the nucleic acid encoding the polypeptide is amplified in lung and colon tumors. Example 143 of the present application provides step-by-step guidelines and protocols for the gene amplification assay. Thus one of ordinary skill in the art would have understood at the time of filing what was encompassed by the claims.

The Examiner asserts that "the skilled artisan cannot envision the *detailed chemical structure* of an encompassed polypeptide, and therefore conception is not achieved until reduction to practice has occurred, regardless of the complexity or simplicity of the method of isolation" (Page 23 of the instant Office Action; emphasis in original). In support of this assertion, the Examiner cites the cases of *Fiers v. Revel* and *Amgen v. Chugai*. (Page 23 of the instant Office Action).

Applicants submit that *Fiers v. Revel* and *Amgen v. Chugai* addressed conception and the written description requirement in the context of DNA-related inventions. The *Amgen* court held that conception of a DNA invention "has not been achieved until reduction to practice has occurred, *i.e.*, until after the gene has been isolated." 927 F.2d 1200 (Fed. Cir.), *cert. denied*, 502 U.S. 856 (1991), at 1206. The *Fiers* court extended this decision into the written description arena, holding that "[i]f a conception of a DNA requires a precise definition, such as by structure, formula, chemical name, or physical properties, as we have held, then a description also requires that degree of specificity." *Fiers*, 984 F.2d at 1171. Since the instant claims are directed to polypeptides, *Fiers* and *Amgen* are distinguished on the facts and do not apply.

More recently, in *Enzo Biochem., Inc. v. Genprobe, Inc.* 296 F.3d 1316 (Fed. Cir. 2002), the court adopted the standard that "the written description requirement can be met by 'showing that the invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics, . . . *i.e.*, complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics." *Id.* at 1324. While the invention in *Enzo* was still a DNA, the holding has been treated as being applicable to proteins as well. Indeed, the court adopted the standard from the USPTO's Written Description Examination Guidelines, which apply to both proteins and nucleic acids.

Accordingly, current applicable case law holds that biological sequences are not adequately described solely by a description of their desired functional activities. The instant claims meet the standard set by the *Enzo* court in that the claimed sequences are defined not only by functional properties, but also by structural limitations. It is well established that a combination of functional and structural features may suffice to describe a claimed genus. "An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, *i.e.*, complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation

between function and structure, or some combination of such characteristics."<sup>3</sup> As discussed above, Applicants have recited structural features, namely, 80% sequence identity to SEQ ID NO:374, which are common to the genus. The genus of claimed polypeptides is further defined by having a specific activity for the encoding nucleic acid, wherein the nucleic acid encoding the polypeptide is amplified in lung and colon tumors. Accordingly, a description of the claimed genus has been achieved.

This particular combination of functional activity and structural homology, as disclosed in the specification, has been recognized by the USPTO as sufficient to describe a claimed genus of polypeptides. The Examiner's attention is respectfully directed to Example 14 of the Synopsis of Application of Written Description Guidelines issued by the U.S. Patent Office, which clearly states that protein variants meet the requirements of 35 U.S.C. §112, first paragraph, as providing adequate written description for the claimed invention even if the specification contemplates but does not exemplify variants of the protein if (1) the procedures for making such variant proteins are routine in the art, (2) the specification provides an assay for detecting the functional activity of the protein and (3) the variant proteins possess the specified functional activity and at least 95% sequence identity to the reference sequence.

As discussed above, the procedures for making the claimed variant polypeptides are well known in the art and described in the specification. The specification also provides an assay, shown in Example 143, for detecting the recited functional activity of the nucleic acids encoding the variant polypeptides. Finally, the claimed variant polypeptides possess both the specified functional activity and a defined degree of sequence identity to the reference sequence, SEQ ID NO:374. Accordingly, the claimed polypeptide variants meet the standards set forth in the Written Description Guidelines.

Thus the specification provides adequate written description for polypeptides having at least 80% identity to SEQ ID NO:374 wherein the nucleic acid encoding the polypeptide is amplified in lung and colon tumors. Applicants therefore respectfully request that the Examiner

---

<sup>3</sup> M.P.E.P. §2163 II(A)(3)(a)

reconsider and withdraw the written description rejection of Claims 28-32 and 39-40 under 35 U.S.C. §112, first paragraph.

**CONCLUSION**

All claims pending in the present application are believed to be in *prima facie* condition for allowance, and an early action to that effect is respectfully solicited.

Please charge any additional fees, including any fees for additional extension of time, or credit overpayment to Deposit Account No. 08-1641, referencing Attorney's Docket No. 39780-2830 P1C38. Please direct any calls in connection with this application to the undersigned at the number provided below.

Respectfully submitted,

Date: November 30, 2005

By: Barrie D. Greene  
Barrie D. Greene (Reg. No. 46,740)

**HELLER EHRMAN LLP**  
275 Middlefield Road  
Menlo Park, California 94025-3506  
Telephone: (650) 324-7000  
Facsimile: (650) 324-0638

SV 2167252 v1

## Correlation between Protein and mRNA Abundance in Yeast

STEVEN P. GYGI, YVAN ROCHON, B. ROBERT FRANZA, AND RUEDI AEBERSOLD\*

*Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730*

Received 5 October 1998/Returned for modification 11 November 1998/Accepted 2 December 1998

We have determined the relationship between mRNA and protein expression levels for selected genes expressed in the yeast *Saccharomyces cerevisiae* growing at mid-log phase. The proteins contained in total yeast cell lysate were separated by high-resolution two-dimensional (2D) gel electrophoresis. Over 150 protein spots were excised and identified by capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS). Protein spots were quantified by metabolic labeling and scintillation counting. Corresponding mRNA levels were calculated from serial analysis of gene expression (SAGE) frequency tables (V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler, *Cell* 88:243-251, 1997). We found that the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data. Indeed, for some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold. Another interesting observation is that codon bias is not a predictor of either protein or mRNA levels. Our results clearly delineate the technical boundaries of current approaches for quantitative analysis of protein expression and reveal that simple deduction from mRNA transcript analysis is insufficient.

The description of the state of a biological system by the quantitative measurement of the system constituents is an essential but largely unexplored area of biology. With recent technical advances including the development of differential display-PCR (21), of cDNA microarray and DNA chip technology (20, 27), and of serial analysis of gene expression (SAGE) (34, 35), it is now feasible to establish global and quantitative mRNA expression profiles of cells and tissues in species for which the sequence of all the genes is known. However, there is emerging evidence which suggests that mRNA expression patterns are necessary but are by themselves insufficient for the quantitative description of biological systems. This evidence includes discoveries of posttranscriptional mechanisms controlling the protein translation rate (15), the half-lives of specific proteins or mRNAs (33), and the intracellular location and molecular association of the protein products of expressed genes (32).

Proteome analysis, defined as the analysis of the protein complement expressed by a genome (26), has been suggested as an approach to the quantitative description of the state of a biological system by the quantitative analysis of protein expression profiles (36). Proteome analysis is conceptually attractive because of its potential to determine properties of biological systems that are not apparent by DNA or mRNA sequence analysis alone. Such properties include the quantity of protein expression, the subcellular location, the state of modification, and the association with ligands, as well as the rate of change with time of such properties. In contrast to the genomes of a number of microorganisms (for a review, see reference 11) and the transcriptome of *Saccharomyces cerevisiae* (35), which have been entirely determined, no proteome map has been completed to date.

The most common implementation of proteome analysis is the combination of two-dimensional gel electrophoresis (2DE)

(isoelectric focusing-sodium dodecyl sulfate [SDS]-polyacrylamide gel electrophoresis) for the separation and quantitation of proteins with analytical methods for their identification. 2DE permits the separation, visualization, and quantitation of thousands of proteins reproducibly on a single gel (18, 24). By itself, 2DE is strictly a descriptive technique. The combination of 2DE with protein analytical techniques has added the possibility of establishing the identities of separated proteins (1, 2) and thus, in combination with quantitative mRNA analysis, of correlating quantitative protein and mRNA expression measurements of selected genes.

The recent introduction of mass spectrometric protein analysis techniques has dramatically enhanced the throughput and sensitivity of protein identification to a level which now permits the large-scale analysis of proteins separated by 2DE. The techniques have reached a level of sensitivity that permits the identification of essentially any protein that is detectable in the gels by conventional protein staining (9, 29). Current protein analytical technology is based on the mass spectrometric generation of peptide fragment patterns that are idiosyncratic for the sequence of a protein. Protein identity is established by correlating such fragment patterns with sequence databases (10, 22, 37). Sophisticated computer software (8) has automated the entire process such that proteins are routinely identified with no human interpretation of peptide fragment patterns.

In this study, we have analyzed the mRNA and protein levels of a group of genes expressed in exponentially growing cells of the yeast *S. cerevisiae*. Protein expression levels were quantified by metabolic labeling of the yeast proteins to a steady state, followed by 2DE and liquid scintillation counting of the selected, separated protein species. Separated proteins were identified by in-gel tryptic digestion of spots with subsequent analysis by microspray liquid chromatography-tandem mass spectrometry (LC-MS/MS) and sequence database searching. The corresponding mRNA transcript levels were calculated from SAGE frequency tables (35).

This study, for the first time, explores a quantitative comparison of mRNA transcript and protein expression levels for a relatively large number of genes expressed in the same metabolic state. The resultant correlation is insufficient for predic-

\* Corresponding author. Mailing address: Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730. Phone: (206) 221-4196. Fax: (206) 685-7301. E-mail: ruedi@u.washington.edu.



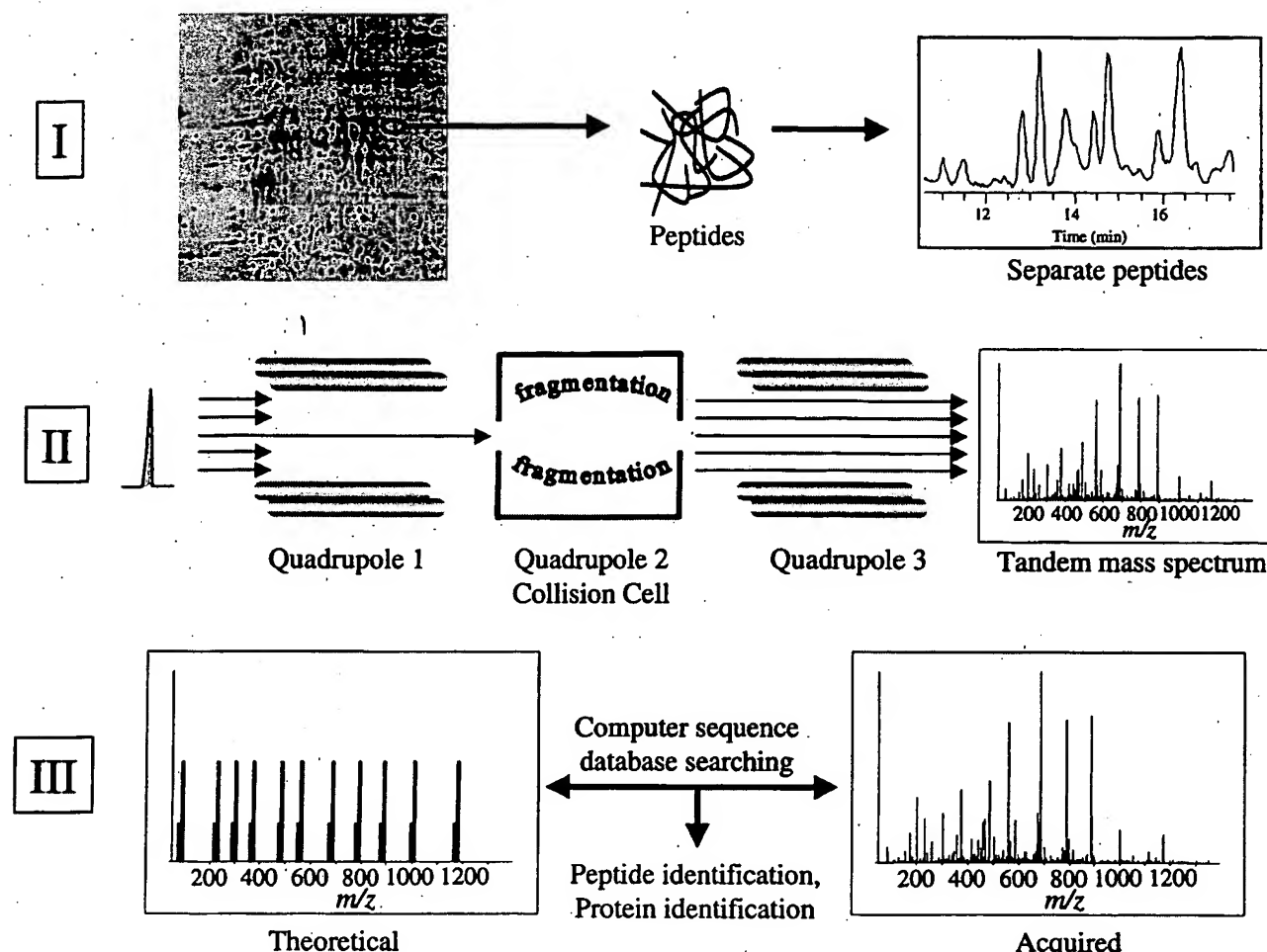


FIG. 1. Schematic illustration of proteome analysis by 2DE and mass spectrometry. In part I, proteins are separated by 2DE, stained spots are excised and subjected to in-gel digestion with trypsin, and the resulting peptides are separated by on-line capillary high-performance liquid chromatography. In part II, a peptide is shown eluting from the column in part I. The peptide is ionized by electrospray ionization and enters the mass spectrometer. The mass of the ionized peptide is detected, and the first quadrupole mass filter allows only the specific mass-to-charge ratio of the selected peptide ion to pass into the collision cell. In the collision cell, the energized, ionized peptides collide with neutral argon gas molecules. Fragmentation of the peptide is essentially random but occurs mainly at the peptide bonds, resulting in smaller peptides of differing lengths (masses). These peptide fragments are detected as a tandem mass (MS/MS) spectrum in the third quadrupole mass filter where two ion series are recorded simultaneously, one each from sequencing inward from the N and C termini of the peptide, respectively. In part III, the MS/MS spectrum from the selected, ionized peptide is compared to predicted tandem mass spectra computer generated from a sequence database. Provided that the peptide sequence exists in the database, the peptide and, by association, the protein from which the peptide was derived can be identified. Unambiguous protein identification is attained in a single analysis because multiple peptides are identified as being derived from the same protein.

tion of protein levels from mRNA transcript levels. We have also compared the relative amounts of protein and mRNA with the respective codon bias values for the corresponding genes. This comparison indicates that codon bias by itself is insufficient to accurately predict either the mRNA or the protein expression levels of a gene. In addition, the results demonstrate that only highly expressed proteins are detectable by 2DE separation of total cell lysates and that therefore the construction of complete proteome maps with current technology will be very challenging, irrespective of the type of organism.

#### MATERIALS AND METHODS

**Yeast strain and growth conditions.** The source of protein and message transcripts for all experiments was YPH499 (*MATa ura3-52 lys2-801 ade2-101 leu2-Δ1 his3-Δ200 trp1-Δ63*) (30). Logarithmically growing cells were obtained by growing yeast cells to early log phase ( $3 \times 10^6$  cells/ml) in YPD rich medium (YPD supplemented with 6 mM uracil, 4.8 mM adenine, and 24 mM tryptophan) at 30°C (35). Metabolic labeling of protein was accomplished in YPD medium

exactly as described elsewhere (4) with the exception that 1 ml of cells was labeled with 3 mCi to offset methionine present in YPD medium. Protein was harvested as described by Garrels and coworkers (12). Harvested protein was lyophilized, resuspended in isoelectric focusing gel rehydration solution, and stored at  $-80^\circ\text{C}$ .

**2DE.** Soluble proteins were run in the first dimension by using a commercial flatbed electrophoresis system (Multiphor II; Pharmacia Biotech). Immobilized polyacrylamide gel (IPG) dry strips with nonlinear pH 3.0 to 10.0 gradients (Amersham-Pharmacia Biotech) were used for the first-dimension separation. Forty micrograms of protein from whole-cell lysates was mixed with IPG strip rehydration buffer (8 M urea, 2% Nonidet P-40, 10 mM dithiothreitol), and 250 to 380  $\mu\text{l}$  of solution was added to individual lanes of an IPG strip rehydration tray (Amersham-Pharmacia Biotech). The strips were allowed to rehydrate at room temperature for 1 h. The samples were run at 300 V–10 mA–5 W for 2 h, then ramped to 3,500 V–10 mA–5 W over a period of 3 h, and then kept at 3,500 V–10 mA–5 W for 15 to 19 h. At the end of the first-dimension run (60 to 70 kV·h), the IPG strips were reequilibrated for 8 min in 2% (wt/vol) dithiothreitol in 2% (wt/vol) SDS–6 M urea–30% (wt/vol) glycerol–0.05 M Tris HCl (pH 6.8) and for 4 min in 2.5% iodoacetamide in 2% (wt/vol) SDS–6 M urea–30% (wt/vol) glycerol–0.05 M Tris HCl (pH 6.8). Following reequilibration, the strips were transferred and apposed to 10% polyacrylamide second-dimension gels. Polyacrylamide gels were poured in a casting stand with 10% acrylamide–2.67% piperazine diacrylamide–0.375 M Tris base–HCl (pH 8.8)–0.1% (wt/vol) SDS–0.05%

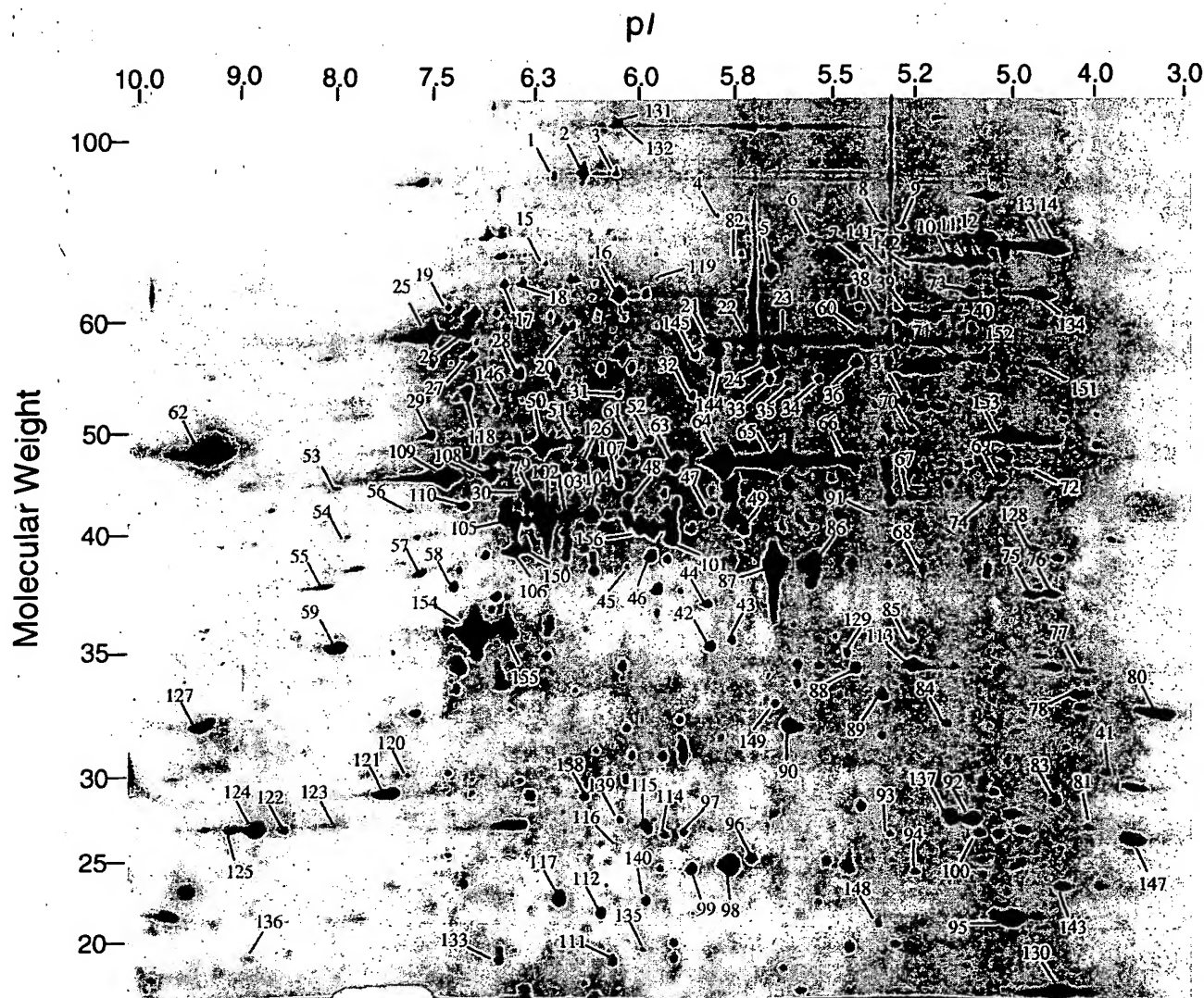


FIG. 2. 2D silver-stained gel of the proteins in yeast total cell lysate. Proteins were separated in the first dimension (horizontal) by isoelectric focusing and then in the second dimension (vertical) by molecular weight sieving. Protein spots (156) were chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. Spots were excised, and the corresponding protein was identified by mass spectrometry and database searching. The spots are labeled on the gel and correspond to the data presented in Table 1. Molecular weights are given in thousands.

(wt/vol) ammonium persulfate–0.05% TEMED (*N,N,N',N'*-tetramethylethylenediamine) in Milli-Q water. The apparatus used to run second-dimension gels was a noncommercial apparatus from Oxford Glycosciences, Inc. Once the IPG strips were apposed to the second-dimension gels, they were immediately run at 50 mA (constant)–500 V–85 W for 20 min, followed by 200 mA (constant)–500 V–85 W until the buffer front line was 10 to 15 mm from the bottom of the gel. Gels were removed and silver stained according to the procedure of Shevchenko et al. (29).

**Protein identification.** Gels were exposed to X-ray film overnight, and then the silver staining and film were used to excise 156 spots of varying intensities, molecular weights, and isoelectric focusing points. In order to increase the detection limit by mass spectrometry, spots were cut out and pooled from up to four identical cold, silver-stained gels. In-gel tryptic digests of pooled spots were performed as described previously (29). Tryptic peptides were analyzed by microcapillary LC-MS with automated switching to MS/MS mode for peptide fragmentation. Spectra were searched against the composite OWL protein sequence database (version 30.2; 250,514 protein sequences) (24a) by using the computer program Sequest (8), which matches theoretical and acquired tandem mass spectra. A protein match was determined by comparing the number of peptides identified and their respective cross-correlation scores. All protein identifications were verified by comparison with theoretical molecular weights and isoelectric points.

**mRNA quantitation.** Velculescu and coworkers have previously generated frequency tables for yeast mRNA transcripts from the same strain grown under the same stated conditions as described herein (35). The SAGE technology is based on two main principles. First, a short sequence tag (15 bp) that contains sufficient information uniquely to identify a transcript is generated. A single tag is usually generated from each mRNA transcript in the cell which corresponds to 15 bp at the 3'-most cutting site for *Nla*III. Second, many transcript tags can be concatenated into a single molecule and then sequenced, revealing the identity of multiple tags simultaneously. Over 20,000 transcripts were sequenced from yeast strain YPH499 growing at mid-log phase on glucose. Assuming the previously derived estimate of 15,000 mRNA molecules per cell (16), this would represent a 1.3-fold coverage even for mRNA molecules present at a single copy per cell and would provide a 72% probability of detecting such transcripts. Computer software which took for input the gene detected, examined the nucleotide sequence, and performed the calculation as described by Velculescu and coworkers (35) was written. In practice, we found that for 21 of 128 (16%) genes examined viable mRNA levels from SAGE data could not be calculated. This was because (i) no CATG site was found in the open reading frame (ORF), (ii) a CATG site was found but the corresponding 10-bp putative SAGE tag was not found in the frequency tables, or (iii) identical putative SAGE tags were present for multiple genes (e.g., TDH2\_YEAST and TDH3\_YEAST).

TABLE 1. Expressed genes identified from 2D gel in Fig. 2

Mol wt	pI	Spot no.	YPD gene name <sup>a</sup>	Protein abundance (10 <sup>3</sup> copies/cell)	mRNA abundance (copies/cell)	Codon bias
17,259	6.75	133	CPR1	15.2	61.7	0.769
18,702	4.80	83	EGD2	20.1	5.2	0.724
18,726	4.44	147	YKL056C	61.2	88.4	0.831
18,978	5.95	135	YER067W	3.7	6.7	0.118
19,108	5.04	130	YLR109W	94.4	9.7	0.680
19,681	9.08	136	ATP7	11.0	NA <sup>b,c</sup>	0.246
20,505	6.07	111	GUK1	16.5	3.7	0.422
21,444	5.25	148	SAR1	5.4	10.4	0.455
21,583	4.98	95	TSA1	110.6	40.1	0.845
22,602	4.30	80	EFB1	66.1	23.8	0.875
23,079	6.29	112	SOD2	12.6	2.2	0.351
23,743	5.44	137	HSP26	NA <sup>d</sup>	0.7	0.434
24,033	5.97	96	ADK1	17.4	16.4	0.656
24,058	4.43	143	YKL117W	29.2	10.4	0.339
24,353	6.30	140	TFS1	8.1	0.7	0.146
24,662	5.85	99	URA5	25.4	6.0	0.359
24,808	6.33	97	GSP1	26.3	5.2	0.735
24,908	8.73	122	RPS5	18.6	NA <sup>c</sup>	0.899
25,081	4.65	81	MRP8	9.3	NA <sup>c</sup>	0.241
25,960	6.06	116	RPE1	5.8	0.7	0.372
26,378	9.55	127	RPS3	96.8	NA <sup>c</sup>	0.863
26,467	5.18	100	VMA4	10.5	3.7	0.427
26,661	5.84	98	TPI1	NA <sup>d</sup>	NA <sup>c</sup>	0.900
27,156	5.56	93	PRE8	6.9	0.7	0.129
27,334	6.13	115	YHR049W	18.4	2.2	0.520
27,472	5.33	92	YNL010W	31.6	3.7	0.421
27,480	8.95	123	GPM1	10.0	169.4	0.902
27,480	8.95	124	GPM1	231.4	169.4	0.902
27,480	8.95	125	GPM1	7.5	169.4	0.902
27,809	5.97	139	HOR2	5.7	0.7	0.381
27,874	4.46	78	YST1	13.6	52.8	0.805
28,595	4.51	41	PUP2	4.4	0.7	0.147
29,156	6.59	114	YMR226C	14.5	2.2	0.283
29,244	8.40	120	DPM1	5.0	11.2	0.362
29,443	5.91	48	PRE4	3.4	3.7	0.162
30,012	6.39	138	PRB1	21.2	1.5	0.449
30,073	4.63	77	BMH1	14.7	28.2	0.454
30,296	7.94	121	OMP2	67.4	41.6	0.499
30,435	6.34	89	GPP1	70.2	11.2	0.703
31,332	5.57	88	ILV6	13.9	3.0	0.402
32,159	5.46	113	IPP1	63.1	3.7	0.752
32,263	6.00	149	HIS1	22.4	4.5	0.232
33,311	5.35	84	SPE3	15.1	6.7	0.468
34,465	5.60	129	ADE1	8.7	5.2	0.305
34,762	5.32	85	SEC14	10.9	6.0	0.373
34,797	5.85	42	URA1	49.5	8.9	0.237
34,799	6.04	90	BEL1	103.2	81.0	0.875
35,556	5.97	43	YDL124W	6.4	4.5	0.206
35,619	8.41	59	TDH1	69.8	32.7 <sup>e</sup>	0.940
35,650	5.49	68	CAR1	5.2	3.0	0.339
35,712	6.72	117	TDH2	49.6	473.0 <sup>e</sup>	0.982
35,712	6.72	154	TDH2	863.5	473.0 <sup>e</sup>	0.982
35,712	6.72	155	TDH2	79.4	473.0 <sup>e</sup>	0.982
36,272	4.85	128	APA1	8.7	0.7	0.425
36,358	5.05	75	YJR105W	17.6	17.1	0.522
36,358	5.05	76	YJR105W	27.5	17.1	0.522
36,596	6.37	79	ADH2	58.9	260.0 <sup>e</sup>	0.711
36,714	6.30	102	ADH1	746.1	260.0	0.913
36,714	6.30	103	ADH1	17.6	260.0	0.913
36,714	6.30	104	ADH1	61.4	260.0	0.913
36,714	6.30	105	ADH1	52.7	260.0	0.913
37,033	6.23	44	TAL1	44.8	3.7	0.701
37,796	7.36	57	IDH2	29.4	6.7	0.330
37,886	6.49	106	ILV5	76.0	4.5	0.892
38,700	7.83	55	BAT1	30.9	11.2	0.469
38,702	6.24	46	QCR2	NA <sup>d</sup>	2.2	0.326

Continued

TABLE 1—Continued

Mol wt	pI	Spot no.	YPD gene name <sup>a</sup>	Protein abundance (10 <sup>3</sup> copies/cell)	mRNA abundance (copies/cell)	Codon bias
39,477	5.58	86	FBA1	17.8	183.6	0.935
39,477	5.58	87	FBA1	427.2	183.6	0.935
39,540	6.50	150	HOM2	60.3	4.5	0.592
39,561	6.12	156	PSA1	96.4	27.5	0.718
41,158	6.01	49	YNL134C	14.9	1.5	0.316
41,623	7.18	58	BAT2	19.0	8.9	0.250
41,728	7.29	110	ERG10	24.1	4.5	0.543
41,900	5.42	74	TOM40	22.3	2.2	0.375
42,402	6.29	45	CYS3	6.7	8.9	0.621
42,883	5.63	67	DYS1	15.8	5.2	0.526
43,409	6.31	107	SER1	10.5	1.5	0.292
43,421	5.59	91	ERG6	2.2	14.1	0.408
44,174	7.32	56	YBR025C	13.1	6.0	0.684
44,682	4.99	72	TIF1	2.9	39.4	0.834
44,707	7.77	108	PGK1	23.7	165.7	0.897
44,707	7.77	109	PGK1	315.2	165.7	0.897
46,080	6.72	30	CAR2	15.4	NA <sup>c</sup>	0.495
46,383	8.52	53	IDP1	7.7	0.7	0.436
46,553	5.98	47	IDP2	32.4	NA <sup>c</sup>	0.197
46,679	6.39	50	ENO1	35.4	0.7	0.930
46,679	6.39	51	ENO1	6.6	0.7	0.930
46,679	6.39	52	ENO1	2.2	0.7	0.930
46,773	5.82	63	ENO2	15.5	289.1	0.960
46,773	5.82	64	ENO2	635.5	289.1	0.960
46,773	5.82	65	ENO2	93.0	289.1	0.960
46,773	5.82	66	ENO2	31.0	289.1	0.960
47,402	6.09	126	COR1	2.5	0.7	0.422
47,666	8.98	54	AAT2	11.7	6.0	0.338
48,364	5.25	73	WTM1	74.5	13.4	0.365
48,530	6.20	61	MET17	38.1	29.0	0.576
48,904	5.18	69	LYS9	16.2	3.7	0.463
48,987	4.90	153	SUP45	29.6	11.9	0.377
49,727	5.47	70	PRO2	13.6	5.2	0.297
49,912	9.27	62	TEF2	558.5	282.0	0.932
50,444	5.67	35	YDR190C	4.8	2.2	0.228
50,837	6.11	32	YEL047C	3.8	1.5	0.387
50,891	4.59	151	TUB2	11.2	7.4	0.404
51,547	6.80	27	LPD1	18.9	2.2	0.351
52,216	7.25	29	SHM2	19.7	7.4	0.722
52,859	5.54	37	YFR044C	30.2	6.7	0.442
53,798	5.19	71	HXX2	26.5	7.4	0.756
53,803	6.05	145	GYP6	4.4	0.7	0.147
54,403	5.29	39	ALD6	37.7	2.2	0.664
54,403	5.29	40	ALD6	6.6	2.2	0.664
54,502	6.20	31	ADE13	6.3	1.5	0.417
54,543	7.75	25	PYK1	225.3	101.8	0.965
54,543	7.75	26	PYK1	39.8	101.8	0.965
55,221	6.66	146	YEL071W	16.3	3.0	0.244
55,295	4.35	134	PDI1	66.2	14.1	0.589
55,364	5.98	24	GLK1	22.6	6.0	0.237
55,481	7.97	118	ATP1	21.6	2.2	0.637
55,886	6.47	28	CYS4	22.2	NA <sup>c</sup>	0.444
56,167	5.83	33	ARO8	14.3	3.0	0.324
56,167	5.83	34	ARO8	9.1	3.0	0.324
56,584	6.36	20	CYB2	18.9	NA <sup>c</sup>	0.259
57,366	5.53	60	FRS2	2.3	0.7	0.451
57,383	5.98	144	ZWF1	5.6	0.7	0.215
57,464	5.49	36	THR4	21.4	3.7	0.508
57,512	5.50	7	SRV2	6.5	NA <sup>c</sup>	0.260
57,727	4.92	152	VMA2	33.7	8.9	0.546
58,573	6.47	17	ACH1	4.4	1.5	0.327
58,573	6.47	18	ACH1	5.4	1.5	0.327
61,353	5.87	21	PDC1	6.5	200.7	0.962
61,353	5.87	22	PDC1	303.2	200.7	0.962
61,353	5.87	23	PDC1	16.3	200.7	0.962
61,649	5.54	38	CCT8	2.2	1.5	0.271

Continued on following page

TABLE 1—Continued

Mol wt	pI	Spot no.	YPD gene name <sup>a</sup>	Protein abundance (10 <sup>3</sup> copies/cell)	mRNA abundance (copies/cell)	Codon bias
61,902	6.21	101	PDC5	4.3	NA <sup>c</sup>	0.828
62,266	6.19	16	ICL1	20.1	NA <sup>c</sup>	0.327
62,862	8.02	19	ILV3	5.3	4.5	0.548
63,082	6.40	119	PGM2	2.2	3.0	0.402
64,335	5.77	5	PAB1	30.4	1.5	0.616
66,120	5.42	8	STI1	6.7	0.7	0.313
66,120	5.42	9	STI1	6.4	0.7	0.313
66,450	5.29	141	SSB2	7.0	NA <sup>c</sup>	0.880
66,450	5.29	142	SSB2	2.3	NA <sup>c</sup>	0.880
66,456	5.23	10	SSB1	64.5	79.5	0.907
66,456	5.23	11	SSB1	59.0	79.5	0.907
66,456	5.23	12	SSB1	13.7	79.5	0.907
68,397	5.82	82	LEU4	3.1	3.0	0.407
69,313	4.90	13	SSA2	24.3	18.6	0.892
69,313	4.90	14	SSA2	77.1	18.6	0.892
74,378	8.46	15	YKL029C	2.8	3.7	0.353
75,396	5.82	6	GRS1	5.5	7.4	0.500
85,720	6.25	1	MET6	2.0	NA <sup>c</sup>	0.772
85,720	6.25	2	MET6	10.9	NA <sup>c</sup>	0.772
85,720	6.25	3	MET6	1.4	NA <sup>c</sup>	0.772
93,276	6.11	131	EFT1	17.9	41.6	0.890
93,276	6.11	132	EFT1	5.7	41.6	0.890
102,064 <sup>c</sup>	6.61 <sup>c</sup>	94	ADE3	4.8	5.2	0.423
107,482 <sup>c</sup>	5.33 <sup>c</sup>	4	MCM3	2.7	NA <sup>c</sup>	0.240

<sup>a</sup> YPD gene names are available from the YPD website (39).<sup>b</sup> NA, calculation could not be performed or was not available.<sup>c</sup> mRNA data inconclusive or NA.<sup>d</sup> No methionines in predicted ORF; therefore, protein concentration was not determined.<sup>e</sup> Measured molecular weight or pI did not match theoretical molecular weight or pI.

**Protein quantitation.** [<sup>35</sup>S]methionine-labeled gels were exposed to X-ray film overnight, and then the silver stain and film were used to excise 156 spots of varying intensities, molecular weights, and pIs. The excised spots were placed in 0.6-ml microcentrifuge tubes, and scintillation cocktail (100  $\mu$ l) was added. The samples were vortexed and counted. In addition, two parallel gels were electroblotted to polyvinylidene difluoride membranes. The membranes were exposed to X-ray film, and four intense single spots were excised from each membrane and subjected to amino acid analysis. For these four spots, a mean of  $209 \pm 4$  cpm/pmol of protein/methionine was found. This number was used to quantitate all remaining spots in conjunction with the number of methionines present in the protein.

To ensure that proteins were labeled to equilibrium, parallel 2D gels were prepared and run on yeast metabolically labeled for 1, 2, 6, or 18 h. The corresponding 156 spots were excised from each gel, and radioactivity was measured by liquid scintillation counting for each spot. Calculated protein levels were highly reproducible for all time points measured after 1 h.

**Calculation of codon bias and predicted half-life.** Codon bias values were extracted from the YPD spreadsheet (17). Protein half-lives were calculated based on the N-end rule (33). When the N-terminal processing was not known experimentally, it was predicted based on the affinity of methionine aminopeptidase (31).

## RESULTS

**Characteristics of proteome approach.** Nearly every facet of proteome analysis hinges on the unambiguous identification of large numbers of expressed proteins in cells. Several techniques have been described previously for the identification of proteins separated by 2DE, including N-terminal and internal sequencing (1, 2), amino acid analysis (38), and more recently mass spectrometry (25). We utilized techniques based on mass spectrometry because they afford the highest levels of sensitivity and provide unambiguous identification. The specific procedure used is schematically illustrated in Fig. 1 and is based on three principles. First, proteins are removed from the gel by

proteolytic in-gel digestion, and the resulting peptides are separated by on-line capillary high-performance liquid chromatography. Second, the eluting peptides are ionized and detected, and the specific peptide ions are selected and fragmented by the mass spectrometer. To achieve this, the mass spectrometer switches between the MS mode (for peptide mass identification) and the MS/MS mode (for peptide characterization and sequencing). Selected peptides are fragmented by a process called collision-induced dissociation (CID) to generate a tandem mass spectrum (MS/MS spectrum) that contains the peptide sequence information. Third, individual CID mass spectra are then compared by computer algorithms to predicted spectra from a sequence database. This results in the identification of the peptide and, by association, the protein(s) in the spot. Unambiguous protein identification is attained in a single analysis by the detection of multiple peptides derived from the same protein.

**Protein identification.** Yeast total cell protein lysate (40  $\mu$ g), metabolically labeled with [<sup>35</sup>S]methionine, was electrophoretically separated by isoelectric focusing in the first dimension and by SDS-10% polyacrylamide gel electrophoresis in the second dimension. Proteins were visualized by silver staining and by autoradiography. Of the more than 1,000 proteins visible by silver staining, 156 spots were excised from the gel and subjected to in-gel tryptic digestion, and the resulting peptides were analyzed and identified by microspray LC-MS/MS techniques as described above. The proteins in this study were all identified automatically by computer software with no human interpretation of mass spectra. They are indicated in Fig. 2 and detailed in Table 1.

The CID spectra shown in Fig. 3 indicate that the quality of the identification data generated was suitable for unambiguous protein identification. The spectra represent the amino acid sequences of tryptic peptides NSGDIVNLGSIAGR (Fig. 3A) and FAVGAFTDSLRL (Fig. 3B). Both peptides were derived from protein S57593 (hypothetical protein YMR226C), which migrated to spot 114 (molecular weight, 29,156; pI, 6.59) in the 2D gel in Fig. 2. Five other peptides from the same analysis were also computer matched to the same protein sequence.

**Protein and mRNA quantitation.** For the 156 genes investigated, the protein expression levels ranged from 2,200 (PGM2) to 863,000 (TDH2/TDH3) copies/cell. The levels of mRNA for each of the genes identified were calculated from SAGE frequency tables (35). These tables contain the mRNA levels for 4,665 genes in yeast strain YPH499 grown to mid-log phase in YPD medium on glucose as a carbon source. In some instances, the mRNA levels could not be calculated for reasons stated in Materials and Methods. For the proteins analyzed in this study, mean transcript levels varied from 0.7 to 473 copies/cell.

**Selection of the sample population for mRNA-protein expression level correlation.** The protein spots selected for identification were selected from spots visible by silver staining in the 2D gel. An attempt was made not to include spots where overlap with other spots was readily apparent. The number of proteins identified was 156 (Table 1). Some proteins migrated to more than one spot (presumably due to differential protein processing or modifications), and protein levels from these spots were calculated by integrating the intensities of the different spots. The 156 protein spots analyzed represented the products of 128 different genes. Genes were excluded from the correlation analysis only if part of the data set was missing; i.e., genes were excluded if (i) no mRNA expression data were available for the protein or putative SAGE tags were ambiguous, (ii) the amino acid sequence did not contain methionine, (iii) more than a single protein was conclusively identified as

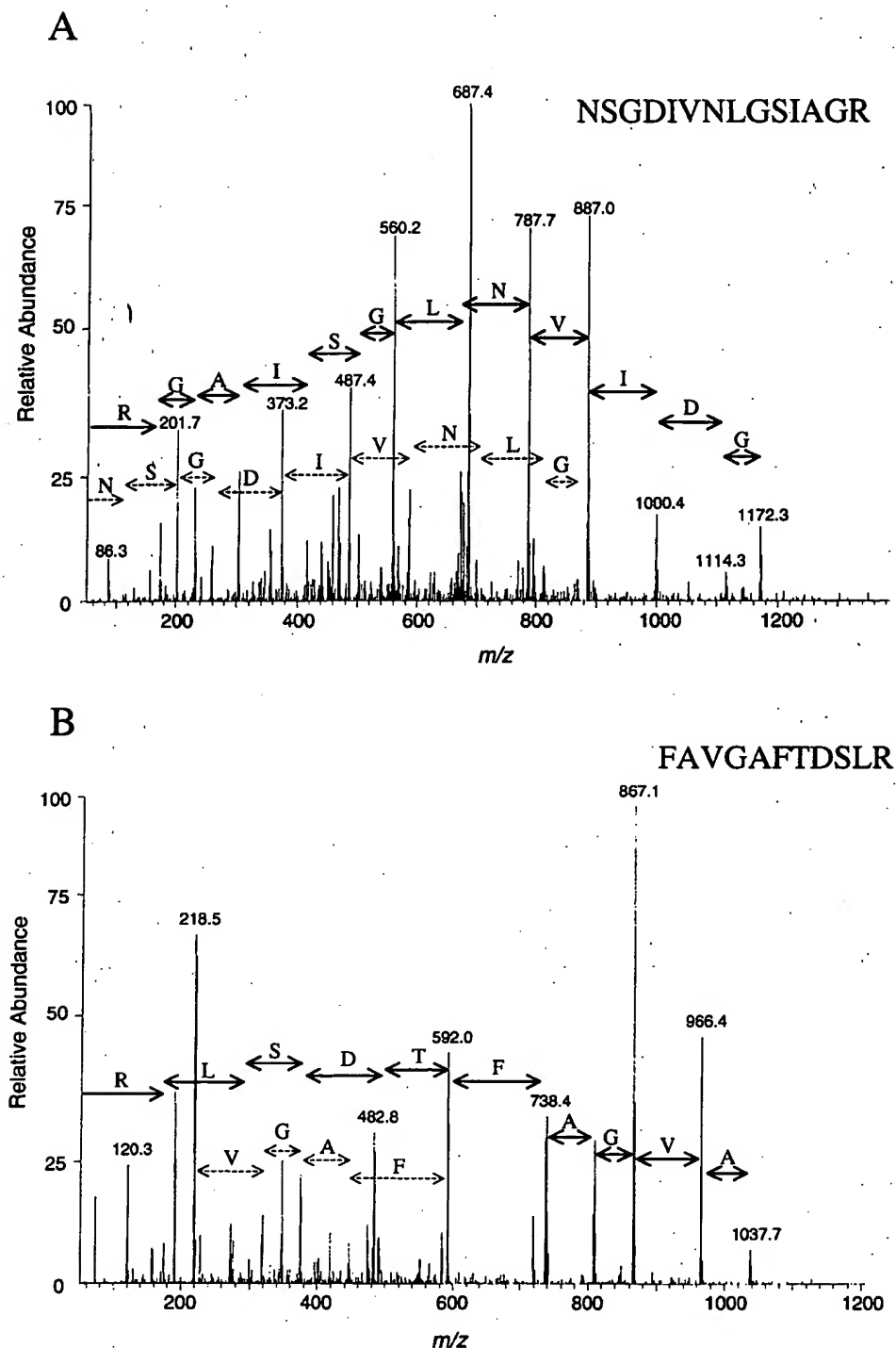


FIG. 3. Tandem mass (MS/MS) spectra resulting from analysis of a single spot on a 2D gel. The first quadrupole selected a single mass-to-charge ratio ( $m/z$ ) of 687.2 (A) or 592.6 (B), while the collision cell was filled with argon gas, and a voltage which caused the peptide to undergo fragmentation by CID was applied. The third quadrupole scanned the mass range from 50 to 1,400  $m/z$ . The computer program Sequest (8) was utilized to match MS/MS spectra to amino acid sequence by database searching. Both spectra matched peptides from the same protein, S57593 (yeast hypothetical protein YMR226C). Five other peptides from the same analysis were matched to the same protein.

migrating to the same gel spot, or (iv) the theoretical and observed pIs and molecular weights could not be reconciled. After these criteria were applied, the number of genes used in the correlation analysis was 106.

**Codon bias and predicted half-lives.** Codon bias is thought to be an indicator of protein expression, with highly expressed proteins having large codon bias values. The codon bias distribution for the entire set of more than 6,000 predicted yeast

gene ORFs is presented in Fig. 4A. The interval with the largest frequency of genes is between the codon bias values of 0.0 and 0.1. This segment contains more than 2,500 genes. The distribution of the codon bias values of the 128 different genes found in this study (all protein spots from Fig. 2) is shown in Fig. 4B, and protein half-lives (predicted from applying the N-end rule [33] to the experimentally determined or predicted protein N termini) are shown in Fig. 4C. No genes were identified with codon bias values less than 0.1 even though thousands of genes exist in this category. In addition, nearly all of the proteins identified had long predicted half-lives (greater than 30 h).

**Correlation of mRNA and protein expression levels.** The correlation between mRNA and protein levels of the genes selected as described above is shown in Fig. 5. For the entire group (106 genes) for which a complete data set was generated, there was a general trend of increased protein levels resulting from increased mRNA levels. The Pearson product moment correlation coefficient for the whole data set (106 genes) was 0.935. This number is highly biased by a small number of genes with very large protein and message levels. A more representative subset of the data is shown in the inset of Fig. 5. It shows genes for which the message level was below 10 copies/cell and includes 69% (73 of 106 genes) of the data used in the study. The Pearson product moment correlation coefficient for this data set was only 0.356. We also found that levels of protein expression coded for by mRNA with comparable abundance varied by as much as 30-fold and that the mRNA levels coding for proteins with comparable expression levels varied by as much as 20-fold.

The distortion of the correlation value induced by the uneven distribution of the data points along the *x* axis is further demonstrated by the analysis in Fig. 6. The 106 samples included in the study were ranked by protein abundance, and the Pearson product moment correlation coefficient was repeatedly calculated after including progressively more, and higher-abundance, proteins in each calculation. The correlation values remained relatively stable in the range of 0.1 to 0.4 if the lowest-expressed 40 to 95 proteins used in this study were included. However, the correlation value steadily climbed by the inclusion of each of the 11 very highly expressed proteins.

**Correlation of protein and mRNA expression levels with codon bias.** Codon bias is the propensity for a gene to utilize the same codon to encode an amino acid even though other codons would insert the identical amino acid in the growing polypeptide sequence. It is further thought that highly expressed proteins have large codon biases (3). To assess the value of codon bias for predicting mRNA and protein levels in exponentially growing yeast cells, we plotted the two experimental sets of data versus the codon bias (Fig. 7). The distribution patterns for both mRNA and protein levels with respect to codon bias were highly similar. There was high variability in the data within the codon bias range of 0.8 to 1.0. Although a large codon bias generally resulted in higher protein and message expression levels, codon bias did not appear to be predictive of either protein levels or mRNA levels in the cell.

## DISCUSSION

The desired end point for the description of a biological system is not the analysis of mRNA transcript levels alone but also the accurate measurement of protein expression levels and their respective activities. Quantitative analysis of global mRNA levels currently is a preferred method for the analysis of the state of cells and tissues (11). Several methods which either provide absolute mRNA abundance (34, 35) or relative

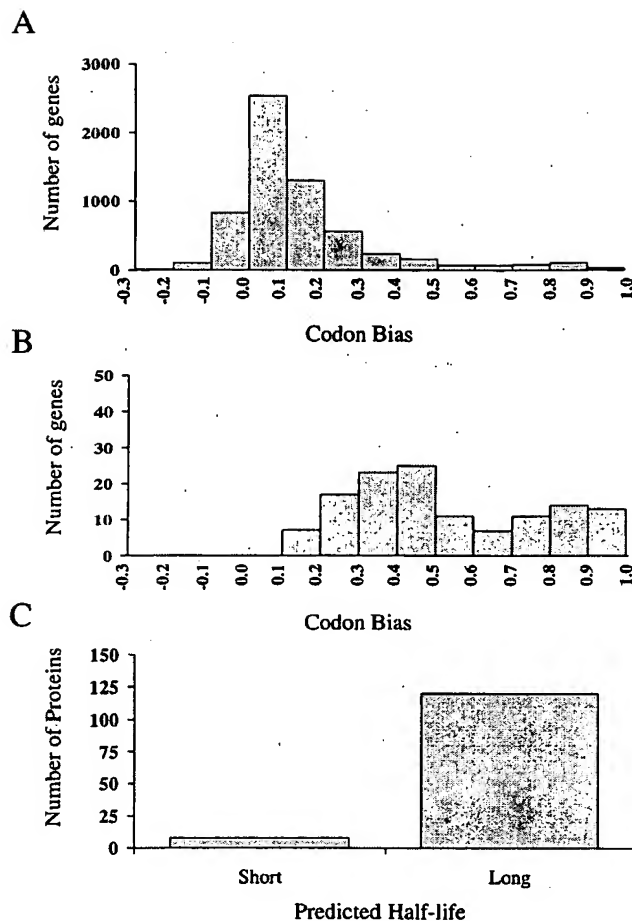


FIG. 4. Current proteome analysis technology utilizing 2DE without pre-enrichment samples mainly highly expressed and long-lived proteins. Genes encoding highly expressed proteins generally have large codon bias values. (A) Distribution of the yeast genome (more than 6,000 genes) based on codon bias. The interval with the largest frequency of genes is 0.0 to 0.1, with more than 2,500 genes. (B) Distribution of the genes from identified proteins in this study based on codon bias. No genes with codon bias values less than 0.1 were detected in this study. (C) Distribution of identified proteins in this study based on predicted half-life (estimated by N-end rule).

mRNA levels in comparative analyses (20, 27) have been described elsewhere. The techniques are fast and exquisitely sensitive and can provide mRNA abundance for potentially any expressed gene. Measured mRNA levels are often implicitly or explicitly extrapolated to indicate the levels of activity of the corresponding protein in the cell. Quantitative analysis of protein expression levels (proteome analysis) is much more time-consuming because proteins are analyzed sequentially one by one and is not general because analyses are limited to the relatively highly expressed proteins. Proteome analysis does, however, provide types of data that are of critical importance for the description of the state of a biological system and that are not readily apparent from the sequence and the level of expression of the mRNA transcript. This study attempts to examine the relationship between mRNA and protein expression levels for a large number of expressed genes in cells representing the same state.

Limits in the sensitivity of current protein analysis technology precluded a completely random sampling of yeast proteins. We therefore based the study on those proteins visible by silver



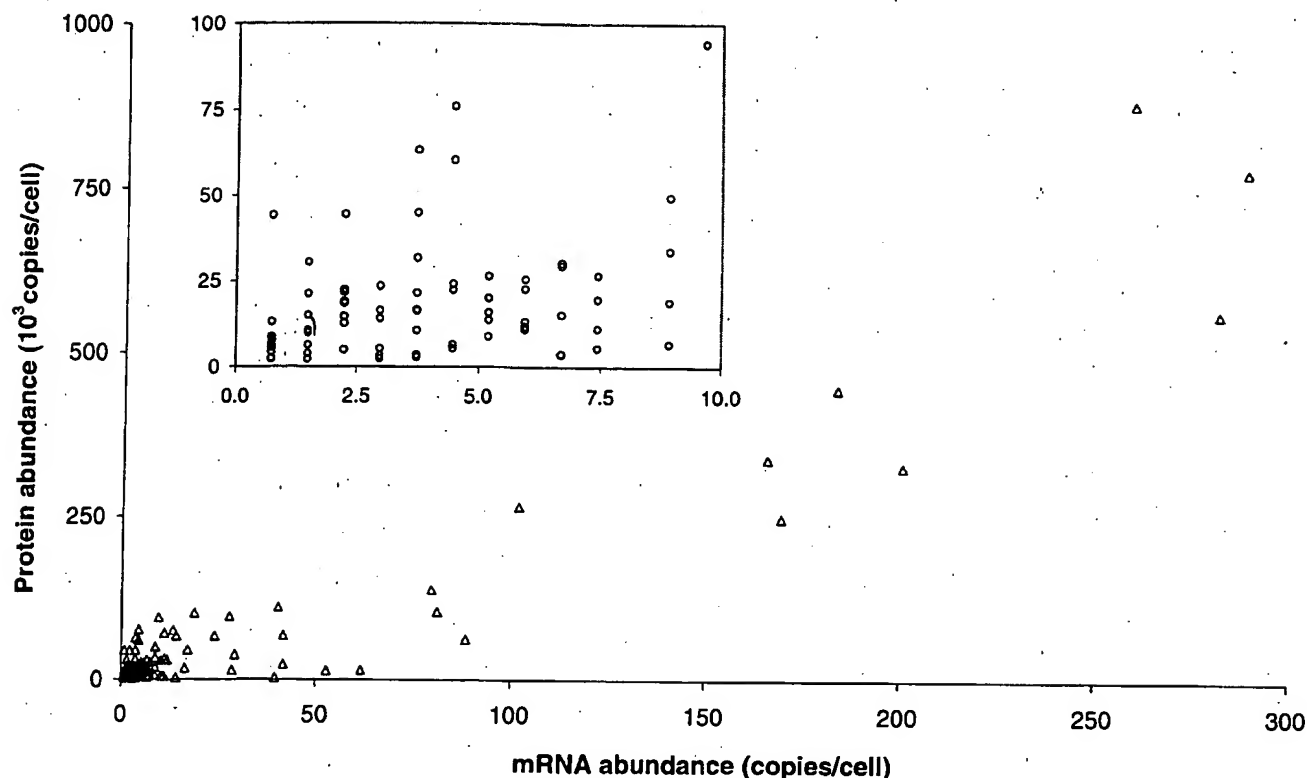


FIG. 5. Correlation between protein and mRNA levels for 106 genes in yeast growing at log phase with glucose as a carbon source. mRNA and protein levels were calculated as described in Materials and Methods. The data represent a population of genes with protein expression levels visible by silver staining on a 2D gel chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. The inset shows the low-end portion of the main figure. It contains 69% of the original data set. The Pearson product moment correlation for the entire data set was 0.935. The correlation for the inset containing 73 proteins (69%) was only 0.356.

staining on a 2D gel. Of the more than 1,000 visible spots, 156 were chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities displayed on the 2D protein pattern. The genes identified in this study shared a number of properties. First, all of the proteins in this study had a codon bias of greater than 0.1 and 93% were greater than 0.2 (Fig. 4B). Second, with few exceptions, the proteins in this study had long predicted half-lives according to the N-end rule (Fig. 4C). Third, low-abundance proteins with regulatory functions such as transcription factors or protein kinases were not identified.

Because the population of proteins used in this study appears to be fairly homogeneous with respect to predicted half-life and codon bias, it might be expected that the correlation of the mRNA and protein expression levels would be stronger for this population than for a random sample of yeast proteins. We tested this assumption by evaluating the correlation value if different subsets of the available data were included in the calculation. The 106 proteins were ranked from lowest to highest protein expression level, and the trend in the correlation value was evaluated by progressively including more of the higher-abundance proteins in the calculation (Fig. 6). The correlation value when only the lower-abundance 40 to 93 proteins were examined was consistently between 0.1 and 0.4. If the 11 most abundant proteins were included, the correlation steadily increased to 0.94. We therefore expect that the correlation for all yeast proteins or for a random selection would be less than 0.4. The observed level of correlation between mRNA and protein expression levels suggests the importance

of posttranslational mechanisms controlling gene expression. Such mechanisms include translational control (15) and control of protein half-life (33). Since these mechanisms are also active in higher eukaryotic cells, we speculate that there is no predictive correlation between steady-state levels of mRNA and those of protein in mammalian cells.

Like other large-scale analyses, the present study has several potential sources of error related to the methods used to determine mRNA and protein expression levels. The mRNA levels were calculated from frequency tables of SAGE data. This method is highly quantitative because it is based on actual sequencing of unique tags from each gene, and the number of times that a tag is represented is proportional to the number of mRNA molecules for a specific gene. This method has some limitations including the following: (i) the magnitude of the error in the measurement of mRNA levels is inversely proportional to the mRNA levels, (ii) SAGE tags from highly similar genes may not be distinguished and therefore are summed, (iii) some SAGE tags are from sequences in the 3' untranslated region of the transcript, (iv) incomplete cleavage at the SAGE tag site by the restriction enzyme can result in two tags representing one mRNA, and (v) some transcripts actually do not generate a SAGE tag (34, 35).

For the SAGE method, the error associated with a value increases with a decreasing number of transcripts per cell. The conclusions drawn from this study are dependent on the quality of the mRNA levels from previously published data (35). Since more than 65% of the mRNA levels included in this study were calculated to 10 copies/cell or less (40% were less

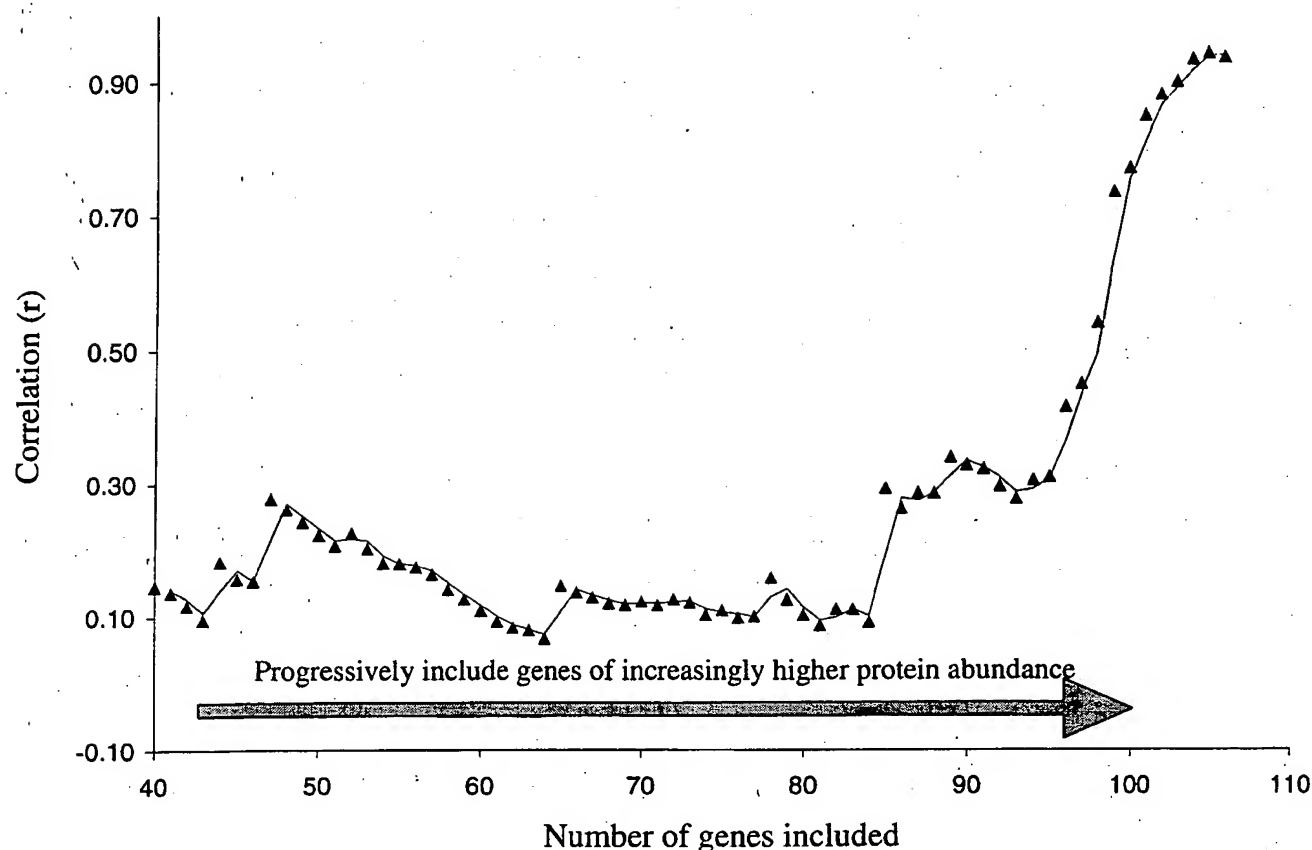


FIG. 6. Effect of highly abundant proteins on Pearson product moment correlation coefficient for mRNA and protein abundance in yeast. The set of 106 genes was ranked according to protein abundance, and the correlation value was calculated by including the 40 lowest-abundance genes and then progressively including the remaining 66 genes in order of abundance. The correlation value climbs as the final 11 highly abundant proteins are included.

than 4 copies/cell), the error associated with these values may be quite large. The mRNA levels were calculated from more than 20,000 transcripts. Assuming that the estimate of 15,000 mRNA molecules per cell is correct (16), this would mean that mRNA transcripts present at only a single copy per cell would be detected 72% of the time (35). The mRNA levels for each gene were carefully scrutinized, and only mRNA levels for which a high degree of confidence existed were included in the correlation value.

Protein abundance was determined by metabolic radiolabeling with [ $^{35}\text{S}$ ]methionine. The calculation required knowledge of three variables: the number of methionines in the mature protein, the radioactivity contained in the protein, and the specific activity of the radiolabel normalized per methionine. The number of methionines per protein was determined from the amino acid sequence of the proteins identified by tandem mass spectrometry. For some proteins, it was not known whether the methionine of the nascent polypeptide was processed away. The N termini of those proteins were predicted based on the specificity of methionine aminopeptidase (31). If the N-terminal processing did not conform to the predicted specificity of processing enzymes, the calculation of the number of methionines would be affected. This discrepancy would affect most the quantitation of a protein with a very low number of methionines. The average number of calculated methionines per protein in this study was 7.2. We therefore expect the potential for erroneous protein quantitation due to unusual N-terminal processing to be small.

The amount of radioactivity contained in a single spot might be the sum of the radioactivity of comigrating proteins. Because protein identification was based on tandem mass spectrometric techniques, comigrating proteins could be identified. However, comigrating proteins were rarely detected in this study, most likely because relatively small amounts of total protein (40  $\mu\text{g}$ ) were initially loaded onto the gels, which resulted in highly focused spots containing generally 1 to 25 ng of protein. Because of the relatively small amount loaded, the concentrations of any potentially comigrating protein would likely be below the limit of detection of the mass spectrometry technique used in this study (1 to 5 ng) and below the limit of visualization by silver staining (1 to 5 ng). In the overwhelming majority of the samples analyzed, numerous peptides from a single protein were detected. It is assumed that any comigrating proteins were at levels too low to be detected and that their influence in the calculation would be small.

The specific activity of the radiolabel was determined by relating the precise amount of protein present in selected spots of a parallel gel, as determined by quantitative amino acid composition analysis, to the number of methionines present in the sequence of those proteins and the radioactivity determined by liquid scintillation counting. It is possible that the resulting number might be influenced by unavoidable losses inherent in the amino acid analysis procedure applied. Because four different proteins were utilized in the calculation and the experiment was done in duplicate, the specific activity calculated is thought to be highly accurate. Indeed, the specific



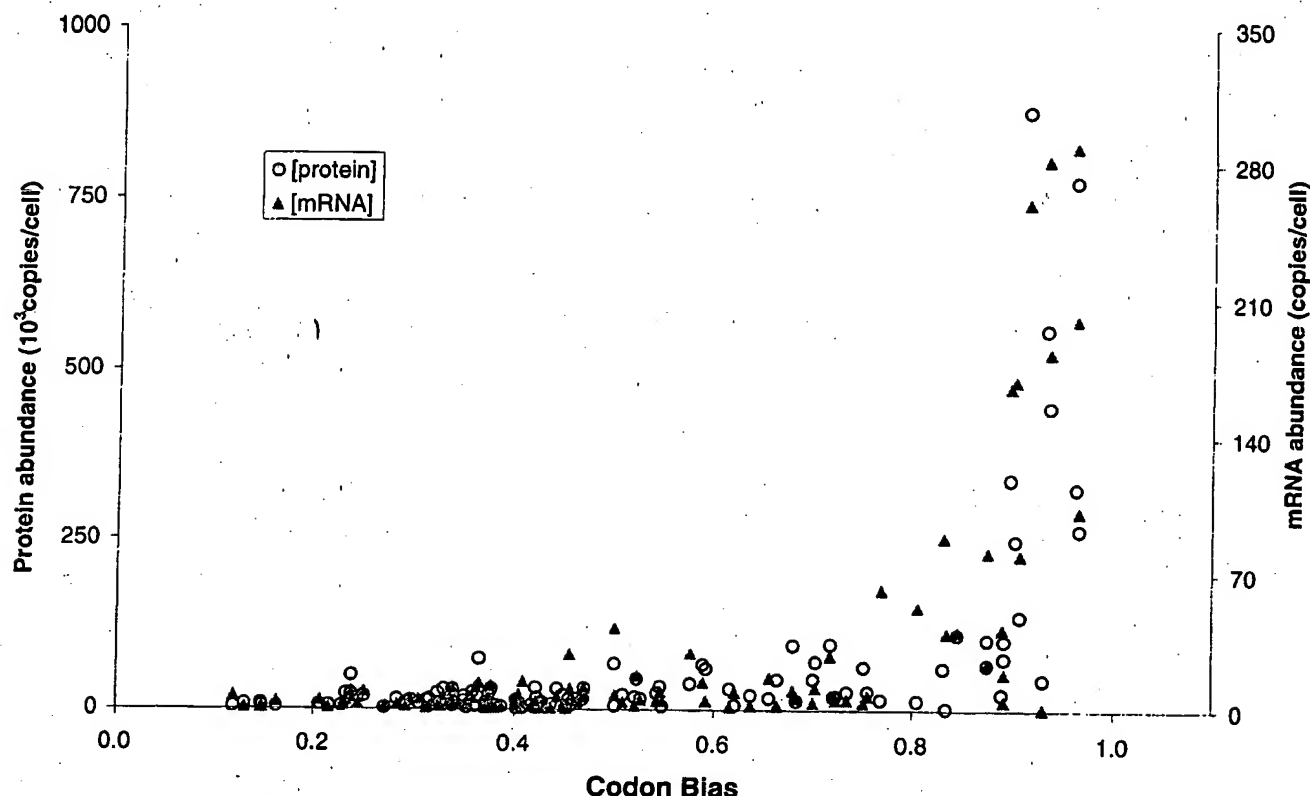


FIG. 7. Relationship between codon bias and protein and mRNA levels in this study. Yeast mRNA and protein expression levels were calculated as described in Materials and Methods. The data represent the same 106 genes as in Fig. 5.

activities calculated for each of the four proteins varied by less than 10%. Any inconsistencies in the calculation of the specific activity would result in differences in the absolute levels calculated but not in the relative numbers and would therefore not influence the correlation value determined.

The protein quantitative method used eliminates a number of potential errors inherent in previous methods for the quantitation of proteins separated by 2DE, such as preferential protein staining and bias caused by inequalities in the number of radiolabeled residues per protein. Any 2D gel-based method of quantitation is complicated by the fact that in some cases the translation products of the same mRNA migrated to different spots. One major reason is posttranslational modification or processing of the protein. Also, artifactual proteolysis during cell lysis and sample preparation can lead to multiple resolved forms of the protein. In such cases, the protein levels of spots coded for by the same mRNA were pooled. In addition, the existence of other spots coded for by the same mRNA that were not analyzed by mass spectrometry or that were below the limit of detection for silver staining cannot be ruled out. However, since this study is based on a class of highly expressed proteins, the presence of undetected minor spots below silver staining sensitivity corresponding to a protein analyzed in the study would generally cause a relatively small error in protein quantitation.

Codon bias is a measure of the propensity of an organism to selectively utilize certain codons which result in the incorporation of the same amino acid residue in a growing polypeptide chain. There are 61 possible codons that code for 20 amino acids. The larger the codon bias value, the smaller the number of codons that are used to encode the protein (19). It is

thought that codon bias is a measure of protein abundance because highly expressed proteins generally have large codon bias values (3, 13).

Nearly all of the most highly expressed proteins had codon bias values of greater than 0.8. However, we detected a number of genes with high codon bias and relative low protein abundance (Fig. 7). For example, the expressed gene with both the second largest protein and mRNA levels in the study was ENO2\_YEAST (775,000 and 289.1 copies/cell, respectively). ENO1\_YEAST was also present in the gel at much lower protein and mRNA levels (44,200 and 0.7 copies/cell, respectively). The codon bias values for ENO2 and ENO1 are similar (0.96 and 0.93, respectively), but the expression of the two genes is differentially regulated. Specifically, ENO1\_YEAST is glucose repressed (6) and was therefore present in low abundance under the conditions used. Other genes with large codon bias values that were not of high protein abundance in the gel include EFT1, TIF1, HXK2, GSP1, EGD2, SHM2, and TAL1. We conclude that merely determining the codon bias of a gene is not sufficient to predict its protein expression level.

Interestingly, codon bias appears to be an excellent indicator of the boundaries of current 2D gel proteome analysis technology. There are thousands of genes with expressed mRNA and likely expressed protein with codon bias values less than 0.1 (Fig. 4A). In this study, we detected none of them, and only a very small percentage of the genes detected in this study had codon bias values between 0.1 and 0.2 (Fig. 4B). Indeed, in every examined yeast proteome study (5, 7, 13, 28) where the combined total number of identified proteins is 300 to 400, this same observation is true. It is expected that for the more complex cells of higher eukaryotic organisms the detection of

low-abundance proteins would be even more challenging than for yeast. This indicates that highly abundant, long-lived proteins are overwhelmingly detected in proteome studies. If proteome analysis is to provide truly meaningful information about cellular processes, it must be able to penetrate to the level of regulatory proteins, including transcription factors and protein kinases. A promising approach is the use of narrow-range focusing gels with immobilized pH gradients (IPG) (23). This would allow for the loading of significantly more protein per pH unit covered and also provide increased resolution of proteins with similar electrophoretic mobilities. A standard pH gradient in an isoelectric focusing gel covers a 7-pH-unit range (pH 3 to 10) over 18 cm. A narrow-range focusing gel might expand the range to 0.5 pH units over 18 cm or more. This could potentially increase by more than 10-fold the number of proteins that can be detected. Clearly, current proteome technology is incapable of analyzing low-abundance regulatory proteins without employing an enrichment method for relatively low-abundance proteins. In conclusion, this study examined the relationship between yeast protein and message levels and revealed that transcript levels provide little predictive value with respect to the extent of protein expression.

#### ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Science and Technology Center for Molecular Biotechnology, NIH grant T32HG00035-3, and a grant from Oxford Glycosciences.

We thank Jimmy Eng for expert computer programming, Garry Corthals and John R. Yates III for critical discussion, and Siavash Mohandesi for expert technical help.

#### REFERENCES

- Aebersold, R. H., J. Leavitt, R. A. Saavedra, L. E. Hood, and S. B. Kent. 1987. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after *in situ* protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* 84:6970-6974.
- Aebersold, R. H., D. B. Teplow, L. E. Hood, and S. B. Kent. 1986. Electrophoretic blotting onto activated glass. High efficiency preparation of proteins from analytical sodium dodecyl sulfate-polyacrylamide gels for direct sequence analysis. *Eur. J. Biochem.* 261:4229-4238.
- Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. *J. Biol. Chem.* 257:3026-3031.
- Boucherie, H., G. Dujardin, M. Kermorgant, C. Monribot, P. Slonimski, and M. Perrot. 1995. Two-dimensional protein map of *Saccharomyces cerevisiae*: construction of a gene-protein index. *Yeast* 11:601-613.
- Boucherie, H., F. Sagliocco, R. Joubert, I. Maillet, J. Labarre, and M. Perrot. 1996. Two-dimensional gel protein database of *Saccharomyces cerevisiae*. *Electrophoresis* 17:1683-1699.
- Carmen, A. A., P. K. Brindle, C. S. Park, and M. J. Holland. 1995. Transcriptional regulation by an upstream repression sequence from the yeast enolase gene *ENO1*. *Yeast* 11:1031-1043.
- Ducret, A., I. VanOostveen, J. K. Eng, J. R. Yates, and R. Aebersold. 1998. High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci.* 7:706-719.
- Eng, J., A. McCormack, and J. R. Yates. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5:976-989.
- Figgeys, D., A. Ducret, J. R. Yates, and R. Aebersold. 1996. Protein identification by solid phase microextraction-capillary zone electrophoresis-micro-electrospray-tandem mass spectrometry. *Nat. Biotechnol.* 14:1579-1583.
- Figgeys, D., I. VanOostveen, A. Ducret, and R. Aebersold. 1996. Protein identification by capillary zone electrophoresis/microelectrospray ionization-tandem mass spectrometry at the subfemtomole level. *Anal. Chem.* 68:1822-1828.
- Fraser, C. M., and R. D. Fleischmann. 1997. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* 18:1207-1216.
- Garrels, J. I., B. Futcher, R. Kobayashi, G. I. Latter, B. Schwender, T. Volpe, J. R. Warner, and C. S. McLaughlin. 1994. Protein identifications for a *Saccharomyces cerevisiae* protein database. *Electrophoresis* 15:1466-1486.
- Garrels, J. I., C. S. McLaughlin, J. R. Warner, B. Futcher, G. I. Latter, R. Kobayashi, B. Schwender, T. Volpe, D. S. Anderson, F. Mesquita-Fuentes, and W. E. Payne. 1997. Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis* 18:1347-1360.
- Gygi, S. P., and R. Aebersold. 1998. Absolute quantitation of 2-DE protein spots, p. 417-421. In A. J. Link (ed.), *2-D protocols for proteome analysis*. Humana Press, Totowa, N.J.
- Harford, J. B., and D. R. Morris. 1997. Post-transcriptional gene regulation. Wiley-Liss, Inc., New York, N.Y.
- Hereford, L. M., and M. Rosbash. 1977. Number and distribution of polyadenylated RNA sequences in yeast. *Cell* 10:453-462.
- Hodges, P. E., W. E. Payne, and J. I. Garrels. 1998. The Yeast Protein Database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 26:68-72.
- Klose, J., and U. Kobalz. 1995. Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis* 16:1034-1059.
- Kurland, C. G. 1991. Codon bias and gene expression. *FEBS Lett.* 285:165-169.
- Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94:13057-13062.
- Liang, P., and A. B. Pardee. 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 257:967-971.
- Link, A. J., L. G. Hays, E. B. Carmack, and J. R. Yates III. 1997. Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* 18:1314-1334.
- Nawrocki, A., M. R. Larsen, A. V. Podtelejnikov, O. N. Jensen, M. Mann, P. Roepstorff, A. Gorg, S. J. Fey, and P. M. Larsen. 1998. Correlation of acidic and basic carrier ampholyte and immobilized pH gradient two-dimensional gel electrophoresis patterns based on mass spectrometric protein identification. *Electrophoresis* 19:1024-1035.
- O'Farrell, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250:4007-4021.
- 24a. OWL Protein Sequence Database. 2 August 1998, posting date. [Online.] <http://bmb5g11.leeds.ac.uk/bmb5dp/owl.html>. [8 January 1999, last date accessed.]
- Patterson, S. D., and R. Aebersold. 1995. Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* 16:1791-1814.
- Pennington, S. R., M. R. Wilkins, D. F. Hochstrasser, and M. J. Dunn. 1997. Proteome analysis: from protein characterization to biological function. *Trends Cell Biol.* 7:168-173.
- Shalon, D., S. J. Smith, and P. O. Brown. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6:639-645.
- Shevchenko, A., O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, H. Boucherie, and M. Mann. 1996. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* 93:14440-14445.
- Shevchenko, A., M. Wilm, O. Vorm, and M. Mann. 1996. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* 68:850-858.
- Sikorski, R. S., and P. Hieter. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* 122:19-27.
- Tsunasawa, S., J. W. Stewart, and F. Sherman. 1985. Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. *J. Biol. Chem.* 260:5382-5391.
- Urlinger, S., K. Kuchler, T. H. Meyer, S. Uebel, and R. Tamp'e. 1997. Intracellular location, complex formation, and function of the transporter associated with antigen processing in yeast. *Eur. J. Biochem.* 245:266-272.
- Varshavsky, A. 1996. The N-end rule: functions, mysteries, uses. *Proc. Natl. Acad. Sci. USA* 93:12142-12149.
- Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. Serial analysis of gene expression. *Science* 270:484-487.
- Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler. 1997. Characterization of the yeast transcriptome. *Cell* 88:243-251.
- Wilkins, M. R., K. L. Williams, R. D. Appel, and D. F. Hochstrasser. 1997. Proteome research: new frontiers in functional genomics. Springer-Verlag, Berlin, Germany.
- Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, and M. Mann. 1996. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* 379:466-469.
- Yan, J. X., M. R. Wilkins, K. Ou, A. A. Gooley, K. L. Williams, J. C. Sanchez, O. Golaz, C. Pasquall, and D. F. Hochstrasser. 1996. Large-scale amino-acid analysis for proteome studies. *J. Chromatogr. A* 736:291-302.
- YPD Website. 6 March 1998, revision date. [Online.] Proteome, Inc. <http://www.proteome.com/YPDhome.html>. [8 January 1999, last date accessed.]

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☒ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**